UNIVERSITÉ DU QUÉBEC À MONTRÉAL

L'ÉMERGENCE DES PRATIQUES DE MODÉLISATION HYBRIDES :DE LEVINS À AUJOURD'HUI

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN PHILOSOPHIE

PAR

JONATHAN ST-ONGE

MAI 2021

REMERCIEMENTS

L'objectif du mémoire est de démontrer l'existence d'un type de modélisation hybride qui se trouve à la croisée du domaine théorique et empirique. On s'intéresse à cette pratique hybride car elle révèle un nouvel environnement où les modèles théoriques s'imprègnent de l'incertitude causée par la complexité du monde qui nous entoure, ou vice versa, à savoir que les modèles empiriques contiennent de plus en plus des suppositions théoriquement informées. Cela dit, il n'y a pas que les modèles qui sont hybrides dans ce mémoire. Le mémoire lui-même est fondamentalement hybride, se trouvant entre le monde des idées propre à la philosophie et celui du domaine naissant des humanités numériques. Si ce carrefour est certes fascinant, il a son lot de défis. Peut-être le plus important est-il de répondre aux attentes distinctes des deux disciplines, dictées par des valeurs et croyances qui sont, dans l'ensemble, assez dissemblables.

Je mentionne cette situation pour souligner à quel point mon entourage fut important pour compléter ma thèse. D'abord, le soutien qui remonte le plus loin est celui de ma mère, qui n'a jamais cessé de croire en moi, de m'encourager et de m'écouter. Ensuite, tout au long de mon périple, il y a eu ma copine et complice, Vanessa. Ensemble, nous avons traversé et surmonté cette épreuve qu'est le mémoire. Elle a su couper dans le texte et reformuler là où je n'osais pas, et pour cela je ne la remercierai jamais assez. J'ai aussi eu la chance d'avoir deux directeurs qui ont su m'épauler dans mon exploration des différents domaines d'intérêts. Comme beaucoup d'étudiant.es, passé.es et futurs, Pierre a été pour moi un allié indéfectible mon parcours. Pour cela, je le remercie sincèrement. Je suis redevable à Christophe pour m'avoir donné l'opportunité de m'initier au domaine des humanités numériques. Il a fait preuve de patience à mes débuts, ce qui m'a permis de persévérer et, éventuellement, de me trouver une niche dans le domaine. Enfin, je tiens à remercier les membres du jury pour leurs commentaires qui ont stimulé ma réflexion et qui m'ont permis d'améliorer considérablement mon mémoire.

Avant-propos

Ce mémoire mène une double vie ; il est à la fois un travail en philosophie et en humanités numériques. Cela s'est avéré être un défi plus important qu'initialement envisagé pour plusieurs raisons. En philosophie, on cherche à définir clairement les phénomènes, les idées et autres objets d'études que l'on pense problématiques et mal compris. Par exemple, on cherche à définir ce que sont les modèles ? Qu'appelle-t-on modèles théoriques et modèles empiriques. Pourquoi est-il légitime de faire des inférences sur le monde à partir de modèles que nous savons être faux ? Différemment, en humanités numériques, du moins dans la branche qui nous intéresse, on se concentre sur les analyses non supervisées, i.e. l'utilisation des méthodes pour la découverte automatique de structures latentes dans un grand volume de données. Par exemple, ces analyses peuvent nous permettre de découvrir des thèmes d'intérêts dans un domaine scientifique particulier.

Le défi de taille du présent travail, et peut-être la source d'une insatisfaction à l'égard des résultats, est de vouloir combiner le travail philosophique avec des méthodes non supervisées. En effet, d'une part, nous nous intéressons à définir qualitativement ce que sont les modèles, en faisant valoir que la taxonomie existante est incomplète, et, d'autre part, nous utilisons des méthodes qui nécessitent un degré d'interprétation important lorsque le moment est venu d'organiser et classer les résultats obtenus. L'enjeu est donc d'éviter la circularité entre les deux parties, c'est-à-dire de trouver une manière d'interpréter les résultats des méthodes non supervisées qui soit indépendante de ce qui est argumenté qualitativement. En d'autres termes, on ne veut pas simplement interpréter les résultats non supervisés à la lumière de ce qui a été argumenté. Idéalement, on voudrait avoir un test quelconque qui permette de vérifier quel modèle—la taxonomie traditionnelle ou le modèle que l'on propose—est favorisé par les résultats obtenus. On doute que le mémoire soit parvenu à pleinement relever ce défi. Sans remettre en cause les apports individuels des travaux philosophiques et des humanités numériques, nous pensons tout de même que ce travail diminue l'écart entre ces deux mondes encore éloignés l'un de l'autre.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Le philosophe des sciences Michael Weisberg propose que la grande diversité des pratiques de modélisation observées en sciences se résume à trois grand types de pratiques : la modélisation physique, mathématique, et computationnelle (Weisberg, 2013). Ce mémoire vise à valider et approfondir la taxonomie de Weisberg en examinant si celle-ci se généralise à un ensemble volumineux d'articles scientifiques. Pour ce faire, on produit une cartographie des pratiques de modélisation à l'aide d'outils computationnels en humanités numériques. Avec cette cartographie, on propose de montrer l'existence d'une pratique hybride irréductible aux trois types de pratiques proposées par Weisberg. Dans la première partie du mémoire, on reprend les travaux de Weisberg et on définit les différentes pratiques de modélisation comme une combinaison de la structure des modèles et de leurs interprétations par les scientifiques. Dans la seconde partie, on présente d'abord le corpus choisi et les techniques utilisées pour réaliser la cartographie, à savoir la modélisation thématique (Blei *et al.*, 2003; Roberts *et al.*, 2019) et le plongement lexical (Bengio *et al.*, 2003; Mikolov *et al.*, 2013), puis les résultats de notre cartographie. Les contributions originales du mémoire sont de démontrer que (i) les *communautés* de topics, ou des sous-réseaux de thèmes latents qui tendent à cooccurrer dans le corpus, nous permettent d'identifier les différents types de modèles, et que (ii) le plongement lexical se révèle un outil diachronique utile pour étudier l'*évolution* des champs sémantiques en lien avec la modélisation. En combinant les deux méthodes, nos résultats suggèrent qu'un type hybride de pratique de modélisation, que nous avions décrit dans le premier chapitre, où nous avions montré qu'elle n'est pas réductible aux types définis par la typologie de Weisberg, semble émerger de la littérature récente. On note toutefois que la démarche utilisée ne parvient pas complètement à représenter l'interprétation des scientifiques, que l'on considère comme un élément important d'une cartographie des pratiques de modélisation. Malgré tout, on soutient que, d'une part, les humanités numériques offrent de nouvelles possibilités pour valider des propositions épistémiques en philosophie des sciences, et, d'autre part, que la philosophie des sciences permet de faire sens de grands volumes de données textuelles caractéristiques de l'ère de l'information. Ce mémoire vise ainsi à mieux comprendre et consolider la relation fructueuse qui émerge entre la philosophie des sciences et les humanités numériques.

Mots-clés : *Pratiques de modélisation — Humanités numériques — Données massives — Modélisation thématique — Plongement lexical*

Keywords : *Model-building — Digital humanities — Big data — Topic network — Word embedding*

INTRODUCTION

Les stratégies de Levins

Richard Levins, à qui on doit le fameux dicton, la « vérité se trouve à la jonction de mensonges indépendants », divisait la modélisation en biologie des populations en trois grandes stratégies (Levins, 1966, p.426). À la base des stratégies, il y a la présence d'un compromis fondamental entre le réalisme, la généralité et la précision des modèles. La première stratégie consiste à construire des modèles généraux et réalistes, au coût de la précision des prédictions. La seconde est de construire des modèles généraux et précis, au détriment du réalisme. La dernière est de sacrifier la généralité en faveur de la précision et du réalisme des modèles.Selon Levins, le compromis vient de la nécessité de synthétiser différentes perspectives—génétique, démographie et même phylogénie—pour expliquer les phénomènes en écologie. Le compromis est le résultat du grand nombre de choix possibles afin de simplifier la complexité qui découle de cette synthèse. Face à cette complexité, Levins défendait la première stratégie, à savoir la construction de modèles simples et réalistes, au détriment de la précision. Cette stratégie, typique de l'école de ce que l'on a appelé « les théoriciens de la simplicité » (Odenbaugh, 2007), s'oppose à la stratégie que Levins appelle de la « force brute », qui est assimilable à la troisième stratégie (modèles précis et réalistes).

Soixante ans plus tard, cette idée est reprise et développée par le philosophe des sciences Michael Weisberg, qui propose à son tour que l'ensemble des pratiques de modélisation en science se divise essentiellement en trois types de modélisation, à savoir la modélisation physique, mathématique, et computationnel (Weisberg, 2003; Weisberg, 2006; Weisberg, 2007; Weisberg et Muldoon, 2009; Weisberg, 2013). Ces pratiques sont caractérisées par le type de modèles du même nom et la manière dont ceux-ci sont in-

terprétés. Au lieu de privilégier l'une des stratégies au détriment des autres, Weisberg montre que, une fois que les intentions et les types d'idéalisations des scientifiques sont inclus dans l'équation, les différentes stratégies sont, en fait, complémentaires. En ce qui concerne la présence d'un compromis fondamental entre les stratégies de modélisation, Weisberg montre que le compromis correspond plutôt aux standards de fidélités qualitatifs et quantitatifs des scientifiques, en fonction du degré de complexité des modèles (Matthewson et Weisberg, 2009; Evans *et al.*, 2013).

Cartographie des modèles

Weisberg distingue les trois types de modèles—physique, mathématique, et computationnel—selon les structures qu'ils devraient représenter (Weisberg, 2013, p.8). Les modèles physiques sont des objets concrets dont les propriétés représentent le phénomène. Ceux-ci comprennent les modèles réduits et les organismes-modèles. Les modèles mathématiques sont des structures mathématiques abstraites dont les propriétés représentent une représentation, elle-même mathématique, du phénomène (bref ce sont des représentations mathématiques de représentations mathématiques). Weisberg note que ce type de modèle a occupé une grande place en philosophie des sciences, notamment en raison de l'approche sémantique des modèles, laquelle met de l'avant ce type de modèles, et notamment des modèles dynamiques décrits par des équations différentielles (Suppes, 1960; Lloyd, 1988; Fraassen, 1980). Les modèles computationnels sont des ensembles de procédures, ou des listes d'instructions, qui représentent le comportement d'un système. Ce type de modèle se rapproche de la seconde stratégie de Levins, où la précision l'emporte sur le reste. Ces trois types de modèles seront qualifiés ici de « théoriques », pour les opposer à une autre forme de modèles, que nous nommerons « empiriques » (ou statistiques).

Bien que l'on retrouve ailleurs d'autres types de modèles que les types proposés par Weisberg, par exemple les modèles verbaux, analogiques, idéalisés, minimaux, explora-

toires, ou encore les modèles de données (Frigg et Hartmann, 2020), ceux-ci ne sont pas des modèles aux yeux de Weisberg, car ils ne correspondent pas à des structures. Ils occupent une autre fonction dans la pratique de la modélisation. Par exemple, tous les modèles sont idéalisés, et parmi les types d'idéalisation, nous trouvons l'idéalisation minimale. Ce faisant, lesdits modèles idéalisés ou minimaux ne correspondent pas à des types de modèles. De même, les modèles sont utilisés de différentes manières. Ils peuvent être exploratoires, ou des outils de compréhension, mais cela ne définit pas le type de modèle. Les modèles verbaux sont mieux compris comme description du modèle, et le modèle de données, qui est une représentation corrigée, rectifiée, ou régimentée des données (Frigg et Hartmann, 2020), est plutôt similaire à la représentation mathématique du phénomène qu'un type de modèle en soi (Weisberg, 2013, p.96).

De plus, Weisberg suggère que les modèles théoriques sont justifiés, ou jugés utiles, dans la mesure où ils ont des propriétés robustes, lesquelles sont censées saisir, au moins partiellement, la structure causale-mécaniste sous-jacente du système d'intérêt. Une conséquence importante de cette approche est que Weisberg ne considère pas non plus les modèles statistiques, c'est-à-dire des modèles empiriques qui décrivent les relations statistiques des données, comme des modèles (théoriques). Étant donné que les modèles statistiques sont principalement utilisés pour faire de bonnes prédictions et pour améliorer la « qualité de l'ajustement » (*goodness of fit*), que les paramètres de ces modèles ne sont pas définis en fonction de quantités biologiques, mais en fonction des données, et qu'ils ne présentent pas de suppositions mécanistes sur les systèmes ciblés, Weisberg fait valoir que l'étude de ceux-ci appartient plutôt à la théorie de la confirmation et à la théorie expérimentale (Weisberg, 2013, p.96).

Pour illustrer cette différence importante, prenons le modèle Lotka-Volterra, que nous utiliserons comme exemple de travail tout au long du premier chapitre, sous une forme statistique et théorique (mathématique). Pour sa forme statistique, on pourrait utiliser un processus autorégressif pour décrire les équilibres prédateurs-proies dans un éco-

système. Un exemple fréquemment utilisé est la relation entre le lièvre et le lynx au Canada entre 1900 et 1920, tel que décrit par le jeu de données de Hewitt (1921). Ce modèle statistique pour séries temporelles suppose que la variable d'intérêt, disons les populations de lièvres H_t, correspond à la fois aux populations de lièvres, H_{t-1}, et de lynx, L_{t-1}, au moment précédent. [1] Ce modèle prend la forme suivante :

$$\mathrm{E}(H_t) = \alpha + \beta_1 H_{t-1} + \beta_2 L_{t-1}$$

Tout comme en régression linéaire simple, les paramètres β contrôlent les différents patrons de séries temporelles possibles, du bruit blanc à la marche aléatoire, et tout ce qui se trouve entre les deux. Si les paramètres sont tels que $\beta_1 < 1$ et $\beta_2 < 1$, les populations tendent à régresser à leurs tailles moyennes, qui est α. Le modèle suppose tout simplement que la population des Lynx dépend d'une combinaison linéaire additive des populations de lynx et lièvre à un temps précédent. Les paramètres du modèle ne sont pas informés par la théorie en ce sens qu'ils servent d'abord à construire une approximation du phénomène afin de mieux prédire la série temporelle d'une variable d'intérêt.

On sait que la relation prédateurs-proies est un cas typique de modélisation mathématique où on a des suppositions *précises*, et *informées par la théorie*, à propos du système latent. Le modèle de Lotka-Volterra décrit les relations prédateurs-proies par les équations différentielles couplées suivantes :

$$\frac{dH}{dt} = b_h H_t - (m_H H_t)\, L_t$$
$$\frac{dL}{dt} = (b_L H_t)\, L_t - m_L L_t$$

1. Cet exemple est inspiré de McElreath (2019, pp. 553-561).

où on suppose que les populations des lièvres à un moment t correspondent à la différence entre leurs taux de naissance, b_H, et de mortalité, m_H, ce dernier dépendant du nombre de lynx. Et les populations de lynx correspondent à la différence entre leurs taux de naissance, b_L, qui dépend de la taille des populations des lièvres et des lynx, et leur taux de mortalité, m_L, qui dépend du nombre de lynx. C'est un modèle déterministe, car une fois que les conditions initiales sont fixées, le système est parfaitement connu. Contrairement au processus autorégressif, les paramètres du modèle Lotka-Volterra sont biologiquement informés. Weisberg ne considère pas le premier modèle comme un objet d'étude de la philosophie de la modélisation théorique, celui-ci faisant partie du domaine empirique, et se consacre exclusivement à l'étude du second type de modèle.

Weisberg soutient que les types de modèles de sa taxonomie sont suffisants pour résumer tous les types de modèles identifiés par la littérature contemporaine en philosophie des sciences, et le bref tour d'horizon que nous avons fait indique certaines des raisons qu'il invoque à cet effet. À un certain niveau, une grande variété de modèles peut être classée parmi la modélisation physique, mathématique ou computationnelle.

Le présent mémoire reprend la taxonomie de Weisberg pour mieux comprendre, d'une part, les pratiques en science de construction de modèles et, d'autre part, pour problématiser ce qu'il advient de la distinction claire voulue entre la modélisation théorique et empirique lorsque les scientifiques commencent à amalgamer les deux. Ainsi, on se demande, est-ce que trois types de modèles sont vraiment suffisants pour résumer la diversité de modèles en science ? Dit autrement, la taxonomie de Weisberg est-elle utile pour mieux comprendre, organiser, voire intervenir sur les manières dont nous pratiquons la science ? Une pratique hybride à la jonction des pratiques théoriques et empiriques ne pourrait-elle pas constituer un autre type de modélisation ? Comment peut-on utiliser des outils en humanités numériques pour valider l'utilité du modèle de Weisberg à concrètement organiser les différentes pratiques de modélisation observées ?

Vers une cartographie à grande échelle des pratiques de modélisation

Les modèles de classification de Weisberg ont l'avantage de se trouver à un niveau épistémique, ce qui a pour effet d'être à la portée d'une enquête empirique. En effet, en délaissant le niveau ontologique, Weisberg avance que sa vision permet véritablement de mieux comprendre et d'organiser la grande variété de pratiques observées en science (Weisberg, 2013, p.20). D'un point de vue méthodologique, Weisberg soutient sa thèse en se basant sur une *lecture rapprochée* de la littérature (Boyd-Graber *et al.*, 2017). Par lecture rapprochée, on entend une lecture critique de textes choisis en fonction de la théorie, notamment celle de la philosophie des sciences. Cette approche, traditionnelle en philosophie des sciences, fait appel à un bagage d'outils critiques qui permet, sur la base de passages clés, d'extrapoler le sens général d'un ensemble de textes de manière logique. Cette approche demande inévitablement un grand investissement de temps et d'efforts.

Dans les dernières années, une approche complémentaire à la lecture rapprochée a émergé sous la bannière des humanités numériques. Par opposition à la lecture rapprochée, les humanités numériques proposent une *lecture distante* qui repose sur des corpus volumineux, et souvent exhaustifs, de textes d'intérêts, analysés à l'aide de techniques computationnelles. Cette approche permet d'étudier des patrons à grande échelle et d'identifier des territoires inexplorés. Les humanités numériques commencent à faire leurs preuves comme outils de confiance en philosophie des sciences, permettant de retrouver des courants d'idées connus (Malaterre *et al.*, 2019), en reconnaissant des tendances émergentes à la croisée de domaines (Hall *et al.*, 2008), et en examinant la manière dont les différents domaines d'étude en sciences sociales utilisent et s'échangent leur langage (McFarland *et al.*, 2013).

Or, les humanités numériques en sont encore à leurs balbutiements et, à ce titre, nécessitent toujours une certaine prudence. Comprendre l'apport d'une lecture distante

aux questionnements philosophiques est un sujet en soi qui mérite un travail de consolidation (Grimmer et Stewart, 2013). Cette thèse s'inscrit dans ce courant. Son objectif est donc double : (i) mettre à l'épreuve la vision de Weisberg avec une cartographie à grande échelle des pratiques de modélisation en science, et (ii) mieux comprendre la contribution des humanités numériques pour répondre aux enjeux traditionnellement philosophiques.

Organisation du mémoire

Le mémoire se divise en deux grandes parties. Dans la première, qui correspond au chapitre 1, on se demande « ce que sont les modèles», et on développe la définition donnée par Weisberg. On examine les justifications de Weisberg pour réinterpréter la modélisation comme une combinaison de structure des modèles et de leurs interprétations. On problématise la division tranchée que fait Weisberg entre la modélisation théorique, à savoir les modèles physiques, mathématiques, et computationnels, et la modélisation empirique, qui comprend la modélisation statistique. Bien que l'on reconnaisse la valeur de cette distinction dans les cas typiques de modélisation mathématique ou statistique, on présente une hypothèse alternative dans laquelle une combinaison de la modélisation mathématique et de la statistique bayésienne produit une pratique hybride irréductible à l'une ou à l'autre. En raison de la structure hybride des modèles dans laquelle on retrouve des suppositions mathématiques et statistiques indissociables l'une de l'autre, en plus d'une interprétation distincte des capacités représentationnelles du modèle pour représenter le monde, on propose de comprendre cette pratique hybride comme complémentaire à la taxonomie de Weisberg.

Dans la deuxième partie, on utilise les humanités numériques pour cartographier à grande échelle les pratiques de modélisation dans le domaine de l'écologie. On y montre que la cartographie met à l'épreuve la catégorisation de Weisberg. On utilise la même logique que dans la sélection de modèles (McElreath, 2020, ch.9), à savoir que la pratique

hybride devient un modèle alternatif avec lequel on tente de « rejeter », ou dans notre cas « étendre », le modèle de Weisberg. Pour ce faire, on cherche d'abord à prouver que les données observées ne sont pas pleinement expliquées par la théorie, puis on démontre que la pratique hybride est complémentaire à la perspective de Weisberg puisqu'elle permet d'expliquer davantage les données que le modèle de Weisberg à lui seul. On cartographie le territoire de modélisation à partir de quatre revues en écologie, à savoir plus de 22,000 articles répartis sur les soixante dernières années. Cette deuxième partie contient trois chapitres. Dans le chapitre 2 (*méthodologie*), on explore le corpus en notre possession, les étapes de nettoyage, et les méthodes utilisées, à savoir la modélisation thématique (Blei *et al.*, 2003; Blei, 2012; Roberts *et al.*, 2019) et le plongement lexical (*word embedding*, mais plus précisément, le modèle de Word2Vec ; Bengio *et al.*, 2003 ; Mikolov *et al.*, 2013). Ensuite au chapitre 3 on présente les *résultats*, et on conclut au dernier chapitre par une *discussion* de la manière dont les résultats d'humanités numériques permettent de répondre aux questionnements philosophiques, et des enjeux actuels de cette approche.

L'apport des humanités numériques en philosophie

Le présent mémoire cherche à confronter les arguments philosophiques de Weisberg aux données en faisant une *cartographie* des pratiques de modélisation en écologie. Cette cartographie consiste à positionner les différentes pratiques de modélisation relativement les unes aux autres à travers différentes revues scientifiques et dans le temps. Avec la cartographie des pratiques de modélisation, on fait le pont entre la philosophie et les humanités numériques. Au lieu de simplement évaluer conceptuellement les modèles de Weisberg, on se questionne pour savoir si celui-ci se généralise sur un ensemble de pratiques de modélisation observé. La cartographie se concentre sur la perspective de Weisberg, car, en plus de s'inspirer fortement des pratiques retrouvées en science, elle a su synthétiser et intégrer les pensées d'auteur.es important.es en philosophie des

sciences, dont Nancy Cartwright, Mary Hesse, Ronald Giere, Patrick Suppes, et bien d'autres.

Nous utilisons deux méthodes computationnelles pour mener à bien notre cartographie à grande échelle, à savoir la modélisation thématique et le plongement lexical. La modélisation thématique permet d'identifier les mots dont le regroupement nous renseigne sur les thèmes latents d'un article, les proportions de ces différents thèmes, leurs relations de codépendance, ainsi que la manière dont ces thèmes évoluent dans le temps. La modélisation thématique permet de vérifier si les constellations de mots qui caractérisent les différentes pratiques de modélisation changent de manière à devenir hybrides, ce que l'on définit comme un enchevêtrement de termes issus de la modélisation théorique (mathématique et computationnelle) et empirique (statistique). Le plongement lexical complémente la modélisation thématique en mettant de côté l'analyse de thèmes pour s'intéresser explicitement aux relations de cooccurrence entre les mots. Pour ce faire, le plongement lexical *enchâsse* (*embed*) les mots dans un espace vectoriel en fonction de leurs différents contextes d'utilisation. Ce type de représentation s'avère particulièrement utile pour évaluer les relations de similarité entre différents mots d'intérêt. En examinant le sens des mots associés aux différentes pratiques de modélisation, ainsi que leur évolution, on peut donc voir si les relations de similarité entre les mots associés aux différentes pratiques tendent vers une hybridation ou restent indépendants.

On compare la contribution des humanités numériques à la philosophie des modèles au processus de validation en apprentissage machine. Ces dernières années, le domaine de l'apprentissage machine a réussi à créer de puissants algorithmes capables d'apprendre par eux-mêmes. Cela dit, les communautés en apprentissage machine sont confrontées à un problème important, celui du *surapprentissage*. Le problème est que, si les algorithmes ont un trop grand degré de liberté, ceux-ci sont capables d'apprendre par coeur les réponses aux tâches de prédiction qui leur sont soumises. Si tel est le cas, les algorithmes sont en mesure de parfaitement prédire l'ensemble de données sur lequel ils

ont appris, mais échouent à généraliser leur apprentissage sur d'autres ensembles de données. Pour s'assurer que l'algorithme évite cette situation, les chercheur.es divisent leurs jeux de données deux, entraîne leurs algorithmes sur l'un, puis valide leurs capacités de généralisation sur l'autre, lequel a été soigneusement caché jusqu'au moment voulu. *Ultimement, cette capacité de généralisation est ce qui compte lorsqu'il vient le temps de démontrer l'apprentissage de l'apprenant.*

De même, on propose d'évaluer la capacité du modèle de Weisberg à représenter les pratiques de modélisation en science sur sa capacité à généraliser sur un ensemble de données non observées. On tient à préciser que cet exercice de cartographie se situe à la croisée de la philosophie et les humanités numériques. Il s'agit de voir comment, d'une part, la philosophie fournit des pistes de recherche pour faire sens d'un grand nombre de données, et, d'autre part, d'évaluer comment les humanités numériques permettent de valider les propositions épistémiques des philosophes à propos des pratiques des scientifiques.

CHAPITRE I

PRÉAMBULE PHILOSOPHIQUE : QU'EST-CE QUE SONT LES MODÈLES ?

Ce premier chapitre est un avant-propos qui sert de *préambule philosophique* à la partie sur les humanités numériques. On prend le temps de bien définir les différentes pratiques de modélisation théorique et empirique pour spécifier nos attentes vis-à-vis une cartographie des pratiques de modélisation. Pour avoir un point d'ancrage, il est nécessaire de clarifier ce que l'on entend par « modélisation théorique »—la modélisation physique, computationnelle et mathématique—et « modélisation empirique », ou modélisation statistique. Cela nous permettra également d'introduire certaines idées qualitatives sur l'existence d'une pratique hybride.

1.1 Les modèles selon Weisberg

1.1.1 Modèles théoriques = structure + interprétation

S'inspirant de Levins, Weisberg soutient qu'il y a trois types de pratiques de modélisation théorique, lesquelles correspondent plus ou moins aux stratégies de Levins. Il y a la modélisation physique, mathématique, et computationnelle. La modélisation mathématique se rapproche le plus de la troisième stratégie (sacrifier la précision pour le réalisme et la généralité), alors que la modélisation computationnelle est similaire sur certains points à la première stratégie (sacrifier la généralité pour la précision et le réalisme), sans toutefois tomber dans l'approche de la «force-brute». Selon Weisberg,

l'approche de la force brute n'est pas équivalente à la modélisation computationnelle, notamment parce que les idéaux représentationnels ne sont pas les mêmes (Weisberg, 2003). Weisberg se distingue de Levins en mettant l'accent sur les intentions des scientifiques en conjonction avec la structure des modèles pour définir les différents types de modélisation.

Selon Weisberg, les types de modélisation sont caractérisés par des types de modèles correspondants, à savoir les modèles physiques, mathématiques, et computationnels. Ces types de modèles, à leur tour, dépendent du type de structure qui est interprété. Ces types de structures ont des pouvoirs explicatifs distincts qui sont liés à ce que les modèles sont censés représenter. Par exemple, bien que Weisberg reconnaisse que les modèles computationnels soient un sous-ensemble des modèles mathématiques, il les considère comme distincts parce que *the procedure itself is the core component of the model, the structure in virtue of which the parts of a target can be explained* (Weisberg, 2013, p.30). De même, les modèles mathématiques sont des structures mathématiques qui représentent les états d'un système latent et des relations entre ces états. On remarque que les modèles physiques constituent une structure distincte, laquelle est absente de chez Levins (et de la plupart des discussions philosophiques sur les modèles).

Bien que Weisberg affirme qu'il n'y a que trois types de modèles, il soutient qu'une grande variabilité au sein des pratiques de modélisation vient du fait que les modèles sont idéalisés et interprétés. Les idéalisations sont des distorsions volontaires introduites dans la manière dont les scientifiques se représentent le monde. Les idéalisations sont typiques des pratiques de modélisation, car il s'agit d'abandonner l'adéquation modèle-monde afin de simplifier le problème ou encore de se limiter à ce que le scientifique pense être vraiment l'essence d'un phénomène. Bien que l'on ne rentre pas dans les détails sur cet aspect de son travail, on mentionne que Weisberg identifie aussi différents types d'idéalisation galiléenne, minimaliste, et de la modélisation multiple et différents idéaux représentationnels (*completeness, simplicity, 1-causal, maxout, et p-generality* ; Weib-

serg, 2013, ch.6). En ce sens, l'élément intentionnel et représentationnel de la construction des modèles chez Weisberg est beaucoup plus fin que les desiderata de Levins.

La notion d'interprétation réfère aux attentes des scientifiques quant à la relation modèle-monde (ce que Weisberg nomme « la relation de dénotation ») et aux normes avec lesquelles les modèles sont évalués. Pour Weisberg, l'interprétation d'un modèle inclut quatre éléments : la description des modèles par les scientifiques, la fonction d'interprétation (*assignment*), la portée attendue du modèle, et les critères de fidélités. On donne un aperçu de ces différents éléments qui composent l'interprétation des modèles.

En premier lieu, une interprétation d'un modèle inclut la description de ce modèle, i.e. l'ensemble des équations, des images, des mots avec lesquels on décrit sa structure. Une première chose à remarquer est que la description est dissociée du modèle lui-même. Bien que la description permet de préciser la nature du modèle, il n'en reste pas moins qu'il s'agit d'une relation plusieurs à plusieurs (*many-to-many*). En effet, un même modèle peut être décrit de plusieurs manières différentes (par une équation, une représentation graphique, en mots) et une même description peut s'appliquer à différents modèles (un ensemble de plans plus ou moins précis interprété de différentes manières) (Weisberg, 2013, pp.34-35). Ces descriptions peuvent être plus ou moins abstraites ou concrètes selon les contextes. Deuxièmement, la fonction d'interprétation spécifie la manière dont les différentes parties du monde sont appliquées sur (*mapped onto*) les parties du modèle. Troisièmement, la portée attendue du modèle spécifie les composantes du modèle conçues comme ayant une valeur représentationnelle. Enfin, Weisberg offre deux critères de fidélité qui concernent l'évaluation de la similarité du modèle avec le monde, à savoir un critère de fidélité dynamique et un critère de fidélité représentationnelle. Le critère de fidélité dynamique spécifie la proximité attendue entre la sortie du modèle (*output*) et l'état du système cible, alors que le critère de fidélité représentationnelle a trait au degré de similarité attendue entre la structure du modèle et la structure causale du système cible pour que le modèle soit perçu comme une représentation adéquate (Weisberg,

2013, p.41).

En résumé, selon Weisberg, les modèles sont des structures idéalisées et interprétées par lesquelles les scientifiques étudient (indirectement) le monde. Les modèles tirent leur pouvoir explicatif de leur capacité à représenter des parties d'un système cible. Il est nécessaire d'inclure les interprétations des scientifiques dans la théorie des pratiques de modélisation, car elles permettent de préciser les critères de fidélité et d'autres informations complémentaires visant à combler les éléments manquants de la structure, en particulier la description du modèle et la portée attendue du modèle.

1.1.2 La relation modèle-monde

Selon Weisberg, les critères de fidélité scientifique déterminent la valeur épistémique de la relation modèle-monde. On s'intéresse à cette relation, car c'est une bifurcation importante où la pratique hybride se distingue des autres pratiques théoriques. Alors que Weisberg suppose que la robustesse et un bon niveau d'ajustement (*fit*) sont suffisants pour évaluer la relation modèle-monde, la pratique hybride reste proche du domaine empirique en mettant l'accent sur la capacité du modèle à valoir pour des données non observées (c'est-à-dire, comme nous verrons ci-dessous, à généraliser son apprentissage sur un ensemble de données non observées).

1.1.3 La notion de similarité

Weisberg défend la notion de similarité modèle-monde pour expliquer comment les scientifiques apprennent sur le monde à travers les modèles. Il suggère que la notion de similarité est relative à l'interprétation, aux connaissances et aux pratiques antérieures du scientifique dans sa communauté de modélisation et de ses objectifs de recherche (Weisberg, 2013, p.135). Plus précisément, Weisberg propose que l'évaluation passe par un jugement de similarité, tel que décrit par les théories en psychologie de la perception,

entre les attributs et mécanismes du modèle et ce qui est observé dans le monde. Ce faisant, la similarité modèle-monde dépendrait de nos intérêts théoriques et serait une question de degrés plutôt qu'en une relation isomorphe entre le modèle et le monde. Comment exactement est-ce que ce jugement opère ? Pour répondre à cette question, Weisberg propose sa théorie des attributs-correspondants pondérés (*weighted feature-matching account*).

La théorie des attributs-correspondants pondérés s'inspire fortement des travaux de Amos Tversky, un pionnier dans le développement des théories sur le jugement et les biais en théorie de la décision, sur la notion de contraste (Tversky, 1977, cité dans Weisberg 2013, p.144). Tversky développe l'idée selon laquelle deux objets dans un ensemble sont similaires en fonction du nombre d'attributs partagés et non partagés. [1] L'essentiel de cette idée se trouve dans l'équation suivante :

$$\mathrm{S(a,b)} = \theta \mathrm{f(A \cap B)} - \alpha \mathrm{f(A - B)} - \beta \mathrm{f(B - A)}$$

où la similarité des objets a et b est égal aux ensembles d'attributs A et B partagés pénalisée par les attributs qu'ils ne partagent pas, le tout étant modulé par une fonction de pondération f et un jeu de paramètre (θ, α, β). Pour Weisberg, les paramètres de pondération représentent l'interprétation relative des agents et les connaissances préalables. À partir de l'équation de Tversky, Weisberg développe une équation plus générale dans laquelle les objets évalués sont les mécanismes et les attributs, lesquels sont définis respectivement par les règles de transition et les états. Au lieu de présenter la formule plus générale, on s'intéresse plutôt à la relation entre le modèle et la représentation mathématique de la cible.

Weisberg soutient que le système cible n'est pas le même que les données empiriques

1. On note que ces idées se retrouvent aussi plus tôt chez Mary Hessy (Hesse, 1966)

collectées. Il affirme que les cibles sont « des ensembles des propriétés du monde réel » (Weisberg, 2013, p.96). Cela dit, dans le cas des modèles mathématiques et computationnels, Weisberg précise que ce sont plutôt les représentations mathématiques de ces propriétés du monde réel qui sont les cibles sur lesquelles le jugement de similarité est établi. Weisberg résume sa pensée en disant que la notion de similarité se base sur comment le modèle est « ajusté » (*fit*) aux données observées, ou du moins aux dimensions qui sont jugées importantes pour expliquer le système cible. L'ajustement est une procédure d'estimation le plus souvent basée sur l'écart entre la prédiction du modèle et les données observées. Cette notion d'ajustement est clé, car cela implique que la notion de similarité se base sur le modèle et ce qui a été observé, à savoir une mesure de rétrodiction. Il représente cette relation d'ajustement entre le modèle et la représentation mathématique de la cible ainsi :

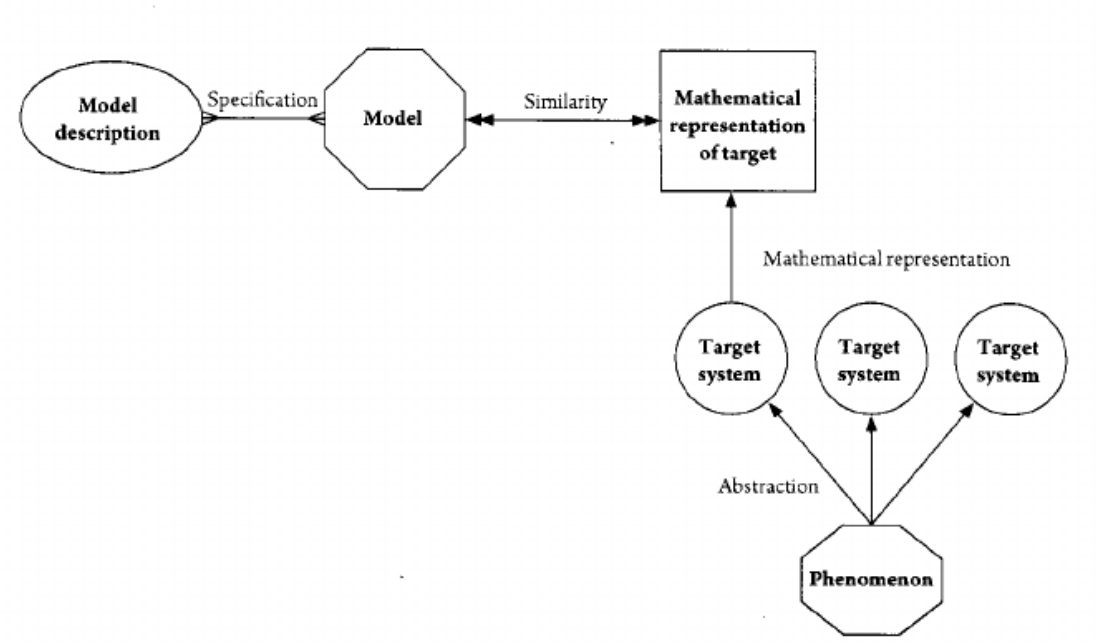


Figure 1.1: La relation modèle-monde pour les modèles mathématiques et computationnels a lieu entre le modèle et la représentation mathématique d'un système cible, lequel représente un phénomène d'intérêt (Weisberg, 2013, p.96)

En résumé, Weisberg affirme que la similarité entre le modèle et le monde est condi-

tionnelle à l'interprétation du scientifique et aux capacités des modèles à représenter les mécanismes et les attributs observés du système cible.

1.1.4 L'analyse de robustesse

Bien que la théorie des attributs correspondants pondérés nous informe sur le jugement du scientifique sur la similarité modèle-monde, elle ne suffit pas à elle seule à expliquer la capacité du modèle à représenter le monde. Pour ce faire, Weisberg reprend et approfondit le thème de la robustesse. L'analyse de robustesse indique que le modèle fait des prédictions fiables lorsque ces prédictions sont semblables à d'autres modèles indépendants (Weisberg, 2006). L'idée originale, popularisée par Levins, est que si différents modèles convergent sur les mêmes prédictions, celles-ci ne dépendent pas du type d'idéalisation dans le modèle, mais réussissent à capturer un ou plusieurs aspects du système cible d'intérêt. Un modèle est robuste s'il continue de représenter le système cible malgré des perturbations.

Selon Weisberg, la robustesse joue un rôle particulier dans les pratiques de modélisation, car elle permet de s'assurer que la similarité entre les modèles et les systèmes cibles est due à la capacité du modèle à représenter le monde, ou la structure causale de celui-ci. Une fois que l'on a en main ces propriétés robustes, on a des raisons de croire que la structure du modèle représente le système cible et qu'il devient possible de contrôler pour l'influence des idéalisations présentes dans le modèle. L'idée étant que si des modèles avec des idéalisations indépendantes prédisent le même phénomène, cela constitue une preuve qu'il existe bel et bien un tel phénomène (que les modèles capturent chacun différemment). Weisberg précise que pour évaluer les prédictions d'un modèle, il faut aussi déterminer si le modèle a une robustesse structurelle et représentationnelle. La robustesse structurelle est la capacité du modèle de rester similaire au système cible malgré des interventions sur sa structure, c'est-à-dire lorsque nous modifions les suppositions ou les paramètres du modèle. De son côté, la robustesse représentationnelle

est la capacité du modèle à posséder les propriétés robustes, malgré le fait que nous changeons son type de représentation.

Comme critère pour établir la qualité de la relation modèle-monde, l'analyse de robustesse soulève un certain nombre de questions. Parmi les plus importantes, on trouve la question de savoir si l'analyse de robustesse, qui ne dépend pas directement de la confrontation avec les données (c.à-d. empirique), peut confirmer si les modèles représentent le monde. [2] Weisberg répond par la négative, mais suggère tout de même que l'analyse de robustesse a un rôle à jouer dans la confirmation. Il écrit qu'une fois qu'il a été établi qu'un modèle est hautement fidèle au système cible, nous pouvons savoir ce qui se passerait si nous intervenions sur ce système (Weisberg, 2013, p.168-9). [3] Weisberg suggère donc que l'établissement de la similarité représentationnelle entre le modèle et le monde est utile, car elle nous permet de généraliser un cas spécifique à d'autres situations au sein du même système.

Dans la prochaine section, on présente la pratique hybride comme un type de modélisation qui possède plusieurs qualités des modèles théoriques (notamment mathématique), mais qui s'inspire davantage des pratiques de modélisation empirique, ou statistique, pour l'établissement du lien modèle-monde. On soutient que l'accent mis sur la prédiction dans la modélisation hybride résulte des communautés de scientifiques qui cherchent à développer des modèles théoriques et complexes capables de faire des prédictions de qualité, notamment sur des phénomènes irréductibles à l'approche expérimentale.

---

2. Bien entendu, l'analyse de robustesse dépend d'une comparaison qualitative avec les données. Par confrontation avec les données, on parle d'utilisation de méthodes statistiques pour quantifier la comparaison.

3. Rappelons que la fidélité d'un modèle est la mesure dans laquelle le modèle doit être similaire au phénomène avant d'être considéré comme une représentation adéquate.

1.2 Les modèles hybrides

On a présenté le modèle de Weisberg et les raisons pour lesquelles celui-ci serait en mesure d'expliquer l'ensemble des pratiques de modélisation en science. Dans cette section, on soutient l'existence d'un autre type de modélisation, hybride qui, en raison de sa structure et de son interprétation distincte, ne se réduit pas à la taxonomie de Weisberg. La modélisation hybride étant à la croisée de la modélisation mathématique et statistique, la preuve de son existence se révèle importante, car celle-ci remet en question les frontières bien établies entre le théorique et l'empirique souvent retrouvées en science.

1.2.1 L'art de la modélisation statistique

Dans son livre *The Art of Statistics*, le statisticien David Spiegelhalter (2019) explique que les statistiques sont un ensemble d'habiletés qui permettent de répondre à des questions d'intérêt à partir de données. De même, le livre de Andrew Gelman et coll. (2020), intitulé *Regression and Other Stories*, propose que les modèles de régression se présentent comme des histoires à raconter, de la collecte des données au processus de validation, en passant par la spécification du modèle statistique et le choix des variables. Dans la même veine, on a un article de James S. Hodges (2019) qui suggère que les «*Statistical methods research [are] done as science rather than mathematics*». Ces trois experts partagent donc l'idée selon laquelle les statistiques ont un statut particulier qui les rapproche davantage d'art ou de la science que des mathématiques.

L'anthropologue et spécialistes des statistiques Richard McElreath utilise quant à lui l'image du modèle géocentrique pour parler des modèles statistiques. Le modèle géocentrique de Ptolémé fut longtemps privilégié parce qu'il faisait des prédictions précises, mais pour ce faire, il utilisait des épicycles, i.e., des cercles étant eux même en rotation sur un cercle plus grand. Avec suffisamment d'épicycles, le modèle géocentrique

est capable de prédire une grande famille de phénomènes. Les modèles statistiques (de régression) sont des approximations linéaires, basées sur une combinaison additive de différentes variables, qui permettent de faire de bonnes prédictions (en moyenne). Au risque d'énoncer une évidence, McElreath suggère que les modèles de régression sont géocentriques lorsque, comme Ptolémé ajoutant des épicycles, les scientifiques ajoutent variable sur variable afin d'extirper des prédictions précises de leurs modèles linéaires.

Par exemple, on se souvient du modèle autorégressif du système proies-prédateurs présenté dans l'introduction :

$$\mathrm{E}\left(H_t\right) = \alpha + \beta_1 H_{t-1}$$

Pour améliorer la prédiction du modèle, on peut ajouter des termes à l'opérateur des retards de manière à ce que la population moyenne de Lièvre corresponde non seulement à la même population à un temps $t-1$ et la population de prédateurs à un même moment, mais aussi, pourquoi pas, à la population de lièvres à un temps $t-2$:

$$\mathrm{E}\left(H_t\right) = \alpha + \beta_1 H_{t-1} + \beta_2 L_{t-1} + \beta_3 H_{t-2}$$

McElreath décrit ce modèle comme géocentrique entre autres parce que les paramètres ne sont pas scientifiquement informés. Il prend comme exemple l'ordonnée à l'origine α, qui dit que même si $H_{t-1} = 0$, il reste possible d'avoir des Lynx à la prochaine période. Cela fait écho à l'idée de Weisberg de rejeter la modélisation statistique sur la base que les modèles théoriques visent l'établissement de la similarité modèle-monde par la haute fidélité entre, d'une part, les mécanismes et les attributs du modèle et, d'autre part, le système cible.

1.2.2 Modèles hybrides = structures hybrides + interprétation

Existe-t-il un intermédiaire entre les modèles mathématiques et statistiques, ou théoriques et empiriques ? Non seulement on argumente que oui, mais on soutient que cet intermédiaire mérite un statut particulier. On réfère aux modèles hybrides comme modèles probabilistes basés sur des principes (MPPs), car ces modèles sont caractérisés par un mélange de suppositions mathématiques et statistiques. [4] On dit que les MPPs sont *probabilistes* et non statistiques parce que les modèles hybrides sont caractérisés par l'utilisation des lois de probabilités pour résoudre un problème essentiel en statistiques inférentielles, à savoir « d'inférer les causes à partir des données » (aussi appelé le problème inverse) [5]

La structure des modèles hybrides

Les modèles hybrides se distinguent des autres types de modèles par une structure qui représente l'interaction entre les états du système, leur évolution (aspect mécaniste des modèles), et leurs différentes sources de variation (aspect probabiliste des modèles). Ce type de modèle est le résultat d'un contexte de modélisation où les scientifiques veulent représenter à la fois le phénomène latent [6] et le processus d'observation, qui inclut le contexte environnant et les outils de mesure utilisés. Pour ce faire, les MPPs exploitent les forces de chaque type de modèles afin d'avoir un modèle dont la structure est scientifiquement informée et capable d'intégrer différentes sources de données dans

---

4. Également appelés «modèles physique-statistique», ou «modèles statistiques motivés par des mécanismes» (Wikle *et al.*, 2019). Richard McElreath (2020) réfère à ces derniers simplement comme des modèles scientifiques ou modèles statistiques ajustés.

5. Voir Jordan 2004 pour une explication de la façon dont on peut interpréter la grande majorité des modèles statistiques comme probabilistes.

6. À noter, on utilise « phénomène » ou « système latent » de manière interchangeable avec l'idée de « système cible ». Dans les deux cas, l'idée est que ce phénomène est non observable et constitue la cible de la modélisation.

un cadre d'analyse quantifiant l'incertitude. Cela permet de développer des modèles théoriques qui représentent l'incertitude du système et qui sont suffisamment complexes pour produire des prédictions de qualité.

La complexité additionnelle requise pour améliorer les capacités prédictives est d'autant plus importante que le système étudié se prête difficilement à un traitement expérimental (ou statistique) dans le style de Ronald Fisher. On qualifie alors ces situations comme étant des contextes empiriquement difficiles. Le cadre expérimental de Fisher est un contexte que l'on pourrait qualifier d'« empiriquement facile », car il y est possible de contrôler les variables environnementales, notamment par le processus de randomisation, pour se prémunir des facteurs confondants. Les partisans des MPPs remettent en question l'idée selon laquelle tous les systèmes peuvent être réduits à ce contexte (Hobbs et Hooten, 2015; Wikle, 2003; Hilborn et Mangel, 1997). Les expériences en écologie sont coûteuses et laborieuses et, comme le mentionne Roughgarden, les phénomènes écologiques sont toujours uniques si nous regardons d'assez près (Roughgarden, 1998). Ainsi, la structure des modèles hybrides vise surtout à représenter des données observationnelles dans lesquelles les effets du système latent et des processus de mesures ne sont pas forcément dissociables. (Patterson *et al.*, 2008; Auger-Méthé *et al.*, 2020).

Le cadre bayésien

Dans la pratique, plusieurs auteurs de la pratique hybride soutiennent que le cadre bayésien est propice à la compréhension des MPPs (Berliner, 1996; Wikle et Hooten, 2010; McElreath, 2020). [7] Un modèle bayésien est essentiellement une interaction entre une fonction de vraisemblance, qui représente la probabilité des observations, étant

7. Dans ce mémoire, on suit essentiellement le cadre bayésien tel que retrouvé chez A. Gelman, M. Betancourt, et R. McElreath. Ces auteurs défendent l'idée selon laquelle le paradigme bayésien est surtout utile pour construire des modèles « ajustés » (*bespoke*), à savoir qui prennent en compte nos connaissances scientifiques du problème, et qui quantifient les différentes sources d'incertitudes.

donné différentes configurations des paramètres du modèle, et des *priors.*[8] Bien que la fonction de vraisemblance abrite le plus souvent un modèle statistique, celle-ci admet aisément une composante mécaniste. La fonction de vraisemblance constitue le modèle de processus et représente la composante mécaniste du phénomène latent d'intérêt. Cela dit, dans la pratique hybride, cette abstraction est indissociable du processus d'observation, du moins pour que le modèle produise de bonnes prédictions.[9]

Les priors sont des distributions de probabilité sur l'espace des paramètres. On utilise les priors pour informer le modèle de l'expertise de domaine sur l'espace de configuration des modèles. On mentionne que l'incorporation des connaissances préalables dans les MPPs est notamment justifiée par la prévention de l'enjeu du surajustement. Une fois que l'on a spécifié le modèle de processus (la fonction de vraisemblance) et les priors, les scientifiques sont en mesure de simuler le processus de génération de données dans son ensemble. Une fois que l'on a en main un modèle capable d'émuler le système cible, on peut ensuite quantifier les configurations qui sont les plus similaires à celui-ci. Ainsi, une autre façon de penser le modèle hybride est comme une combinaison de modèles déterministes, écrite à l'aide de fonctions scientifiquement informées, pour décrire les mécanismes avec un modèle empirique, écrit en utilisant le langage des probabilités, pour décrire les relations dans les données.

---

8. Certains auteurs traduisent le terme « prior » par « modèle de connaissances préalables », ou « croyances antérieures ». Suivant Gelman *et al.*, 2013, on ne pense pas que les priors représentent nécessairement les croyances du modélisateur, et donc que croyance antérieure n'est pas une bonne traduction. Pour alléger le texte et souligner la subtilité du terme, nous utilisons le terme anglais « prior » pour le reste du texte.

9. Le modèle de processus est indissociable du processus d'observation à différent degré. Plus tard, nous verrons qu'un contexte expérimental permet quand même de réduire la structure du processus d'observation. Cela dit, les partisans de la pratique hybride soulignent le fait que ce n'est pas tous les systèmes qui sont réductibles au cadre expérimental, et donc qu'il soit nécessaire de modéliser l'intégralité du processus de génération de donnée.

Un modèle hybride de Lotka-Volterra

On obtient une représentation hybride du modèle Lotka-Volterra en enchâssant celui-ci dans un modèle hiérarchique bayésien. Un exemple d'une source de variation dans le jeu de données de Hewitt est que la relation lièvre-lynx ne contient pas véritablement le compte de la population à l'étude. Il s'agit d'un échantillon de la traite des fourrures de lièvres et de lynx qui varie d'une année à l'autre et dont les chiffres ont été arrondis à la centaine près, puis divisé par 1000(McElreath, 2020). Ainsi, McElreath propose dans son modèle hybride de Lotka-Volterra que le taux de capture des lièvres varie selon la Loi bêta $p_t \sim \text{Beta}(2, 18)$. Pour une population de 10,000 lièvres, cela implique en moyenne 1000 lièvres capturés (une moyenne de 10%). La capture du lièvre comme telle provient d'une Loi binomiale avec la probabilité p_t(rappelons-nous, la loi binomiale représente des événements binaires), laquelle est arrondie à la centaine la plus près et divisée par 1000, on obtient la distribution suivante :

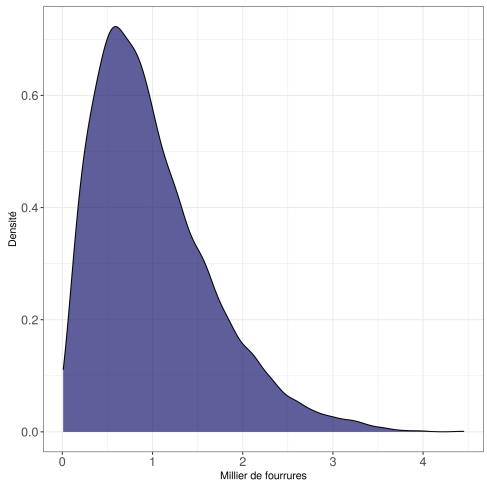


Figure 1.2: La probabilité d'attraper un lièvre est donnée par la loi binomiale avec une probabilité de succès elle-même tirée de la loi bêta.

Contrairement à la modélisation mathématique, la pratique hybride se caractérise donc par la modélisation du processus de génération de données, lequel comprend la dy-

namique des populations et le processus d'observation (McElreath, 2020, p. 557). Ce changement de cible est particulièrement important pour comprendre l'interprétation distincte de la relation modèle-monde dans le cadre de la pratique hybride (Sec. 1.2.3). De manière analogue à la représentation computationnelle de Lotka-Volterra (Weisberg et Reisman, 2008), on propose que sa représentation hybride suivante soit un type distinct de structure (Figure 1.3).

$$
\begin{aligned}
h_t &\sim \text{Log-Normal}(\log(p_H H_t), \sigma_H) && \text{[Prob observed hare pelts]} \\
\ell_t &\sim \text{Log-Normal}(\log(p_L L_t), \sigma_L) && \text{[Prob observed lynx pelts]} \\
H_1 &\sim \text{Log-Normal}(\log 10, 1) && \text{[Prior for initial hare population]} \\
L_1 &\sim \text{Log-Normal}(\log 10, 1) && \text{[Prior for initial lynx population]} \\
H_{T>1} &= H_1 + \int_1^T H_t(b_H - m_H L_t)\mathrm{d}t && \text{[Model for hare population]} \\
L_{T>1} &= L_1 + \int_1^T L_t(b_L H_t - m_L)\mathrm{d}t && \text{[Model for lynx population]} \\
\sigma_H &\sim \text{Exponential}(1) && \text{[Prior for measurement dispersion]} \\
\sigma_L &\sim \text{Exponential}(1) && \text{[Prior for measurement dispersion]} \\
p_H &\sim \text{Beta}(\alpha_H, \beta_H) && \text{[Prior for hare trap probability]} \\
p_L &\sim \text{Beta}(\alpha_L, \beta_L) && \text{[Prior for lynx trap probability]} \\
b_H &\sim \text{Half-Normal}(1, 0.5) && \text{[Prior hare birth rate]} \\
b_L &\sim \text{Half-Normal}(0.05, 0.05) && \text{[Prior lynx birth rate]} \\
m_H &\sim \text{Half-Normal}(0.05, 0.05) && \text{[Prior hare mortality rate]} \\
m_L &\sim \text{Half-Normal}(1, 0.5) && \text{[Prior lynx mortality rate]}
\end{aligned}
$$

Figure 1.3: Modèle hybride de Lotka-Volterra (tiré de McElreath (2020), pp. 546-447)

Dans la Figure 1.3, on voit que la quantification des connaissances préalables sur les relations de données passe par les priors sur les probabilités d'observer les fourrures, les populations initiales, la probabilité de capture, le taux de naissance et mortalité des proies-prédateurs. Si ces priors sont justes, ils permettent de significativement améliorer la capacité du modèle à généraliser des prédictions sur des ensembles de données non

observées, et du même coup valident que le modèle de processus aussi a su capturer le processus latent (Gelman *et al.*, 2017).

Les MPPs ont donc une double identité. Ils démontrent des vertus propres aux modèles théoriques tels que décrits par Weisberg, à savoir que les paramètres des modèles sont informés par la théorie, qu'il y a un usage stratégique des idéalisations, et dont la structure est spécifiée par l'interprétation. En revanche, les MPPs possèdent certaines propriétés que l'on sait être typiques des statistiques, comme la quantification de l'incertitude et un accent mis sur les types de données et leur provenance. On note que la cible n'est plus la même que chez Weisberg, ce qui bouleverse les critères de fidélité proposés par Weisberg. Notamment, les critères de fidélités dans la pratique hybride ne portent plus seulement sur les capacités représentationnelles de la structure, mais sur la capacité du modèle à résoudre le problème inverse, i.e., déterminer les causes de nos observations (sachant que différentes causes peuvent les avoir générées).

La cible est le processus observationnel

Dans la pratique hybride, la cible de la modélisation est le processus observationnel (voir Figure 1.4). Contrairement à la cible que Weisberg propose, le processus observationnel comprend non seulement la représentation mathématique du phénomène d'intérêt, mais aussi le contexte environnant, et la sonde (qui comprend la collecte et le nettoyage de donnée, et la transformation des données existantes) avec laquelle le scientifique explore l'environnement. Plus précisément, le processus observationnel est défini comme le système latent qui engendre les observations à partir desquelles le modèle apprend, ce qui est particulier à la pratique hybride (Betancourt 2019). On présente d'abord cette nouvelle cible qui est représentée par la structure hybride, puis on revient plus loin sur l'apprentissage dans la section sur la relation modèle-monde (Sec. 1.2.3).

Comme c'est le cas dans les autres pratiques de modélisation, les MPPs prennent pour

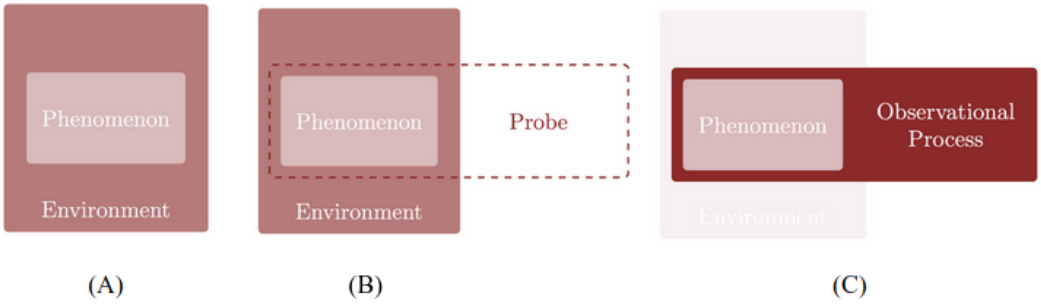


Figure 1.4: Le processus observationnel est une combinaison du phénomène d'intérêt, de l'environnement, et de la sonde. (A) Le phénomène d'intérêt n'existe jamais en isolation. (B) Lorsque nous interagissons avec le phénomène et son contexte, c'est toujours par une sonde observationnelle. (C) Cette combinaison de phénomène, du contexte, et de la sonde compose le processus observationnel, qui est la nouvelle cible de modélisation pour la pratique hybride. Tiré de (Betancourt, 2019)

cible une représentation mathématique du monde. Bien que la cible soit le processus observationnel, les partisans de la pratique hybride interprètent cette cible comme une abstraction (Betancourt, 2019). Or, selon Michael Betancourt, physicien et l'un des développeurs principaux du langage de programmation probabiliste STAN, cet espace représentationnel ne représente pas seulement le phénomène. Contrairement aux autres pratiques de modélisation, la représentation mathématique est mieux comprise comme une variation de différentes possibilités sur un espace d'observation (Betancourt, 2018). Cette variation dans la manière dont les observations sont réalisées peut prendre plusieurs formes.

La pratique hybride distingue la variation ontologique, ou la variation qui est intrinsèque au système latent, de la variation épistémique, ou celle qui appartient à la capacité de résolution de nos instruments de mesure (pris au sens large). La variation ontologique, aussi appelée l'aléatoire aléatorique, représente la variabilité inhérente à chaque fois que « l'on répète l'expérience », alors que l'incertitude épistémique est causée par notre ignorance et les limites physiques de nos instruments de mesure. Betancourt suggère qu'un exemple intéressant de variabilité épistémique se trouve dans la théorie du chaos, dans laquelle de minuscules variations dans les conditions initiales provoquent une dynamique

qui bien qu'imprévisible pour nous, est déterministe (Betancourt, 2019). [10]

Une fois que la manière dont les observations sont réalisées dans le processus observationnel est spécifiée, on obtient le processus de génération de données. En effet, celui-ci est la représentation mathématique du système cible dans la modélisation hybride (Betancourt, 2019). Étant donné que cet espace est le résultat de la variation de différentes possibilités dans un espace d'observation, les modèles hybrides représentent le processus de génération de données à l'aide de probabilités de distribution.

On soutient que cette interprétation de la cible en tant que processus génératif sous-jacent aux données, ce que les partisans de la pratique hybride ont appelé la pensée générative, est une nouvelle pratique de modélisation (McElreath, 2020; Gelman *et al.*, 2013). La pensée générative est une stratégie de modélisation dans laquelle les scientifiques utilisent un même modèle pour simuler des observations et estimer les paramètres, puis complexifient progressivement le modèle pour représenter la sonde et le système latent. L'idée est que, au moins dans certains contextes, les relations avec les phénomènes d'intérêt sont limitées par le processus de mesure et que le mieux que l'on puisse faire est de les modéliser comme un processus de génération de données.

1.2.3 La relation modèle-monde dans la pratique hybride

Le lien de similarité passe par l'actualisation bayésienne

Avec la pratique hybride, on a un mélange de suppositions mécanistes et probabilistes qui définit un espace de configuration du modèle, où chaque configuration est une structure générative qui est un récit mathématique de la façon dont les données auraient pu être générées (Betancourt, 2019). Une fois que le modèle de processus et les priors

10. On note que Betancourt lui-même reconnaît que cet exemple est sujet à discussion. Cette incertitude entre ce qu'est la variation ontologique et épistémique est au cœur du cadre bayésien, et l'une des raisons pour lesquelles la pratique hybride favorise celui-ci.

sont spécifiés, les scientifiques utilisent le théorème de Bayes (ou plutôt une approximation de celui-ci) pour actualiser l'espace préalable en une distribution postérieure (voir Figure 1.5). On mentionne que cette actualisation des croyances du modèle par les données est également comprise comme un processus d'apprentissage. Ainsi, les modèles sont aussi des structures qui apprennent dans la mesure où elles permettent d'élaguer des configurations invraisemblables selon les suppositions présentes dans la structure. Ce processus d'apprentissage permet de quantifier les configurations du modèle qui sont les plus compatibles avec notre expertise de domaine et le processus de génération de données (Gabry *et al.*, 2019; Betancourt, 2019). La similarité modèle-monde porte sur cette compatibilité entre nos expertises de domaine et le processus de génération de données.

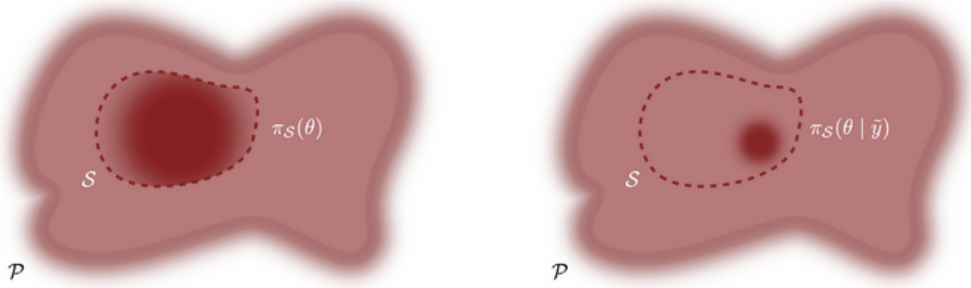


Figure 1.5: L'espace de configuration du modèle (le « petit monde », S), une fois que celui-ci est quantifié par une distribution de probabilité. À gauche, l'espace n'est pas encore informé par les données, et donc les configurations sont dominées par les priors, $\pi_S(\theta)$. À droite, l'espace de configuration du modèle après que le modèle ait actualisé ses croyances en fonction des nouvelles évidences à l'aide du théorème de Bayes, $\pi_S(\theta|\tilde{y}) \propto \pi_S(\tilde{y}|\theta) \cdot \pi_S\theta)$. Pour distinguer le modèle qui n'est pas évalué, de celui qui est spécifié par une fonction de densité, Betancourt dénote une observation $\tilde{y}$. Tiré de Betancourt (2019)

L'interaction entre le modèle de processus et le modèle d'observations fournit une distribution postérieure qui ne peut être réduite à l'un ou à l'autre isolément (Gelman *et al.*, 2013; Gelman *et al.*, 2017). Ceci a pour effet que la pratique hybride ne conçoit pas la stochasticité comme une « couche de bruit » indépendante qui brouille le signal du

«vrai» système latent.[11] La pratique hybride hérite du paradigme bayésien l'idée que ce qui est connu et ce qui est inconnu, ou l'observable et l'inobservable, sont susceptibles de changer, et que parfois la distinction entre les deux n'est pas aussi claire que l'on souhaite. La nature de certains systèmes a pour effet que l'on ne peut pas différencier les différents types de variations, à savoir ontologique et épistémique (Betancourt, 2018; Hilborn et Mangel, 1997). En conséquence, l'approche hybride conserve l'interaction entre les processus latents et processus d'observation pour résoudre le problème inverse.

La précision prédictive attendue pour s'assurer que la similarité n'est pas illusoire

Puisque l'approche hybride conçoit davantage les modèles hybrides comme des outils pour résoudre le problème inverse, elle met de l'avant la précision prédictive attendue pour s'assurer que la similarité entre le modèle-monde n'est pas illusoire (en plus de l'analyse de robustesse). Comme son nom l'indique, la précision prédictive attendue est une mesure qui quantifie la capacité du modèle à prédire ce qui n'est pas observé, sans utiliser de nouvelles observations. Il s'agit d'une mesure pour estimer la capacité de généralisation d'apprentissage du modèle. Cela ne signifie pas qu'elle rejette la nécessité d'établir une haute-fidélité entre le modèle et le monde, mais que cette fidélité doit être complétée par la capacité de la structure à généraliser sur des situations non observées. À la différence du modèle de Weisberg, qui place tout le poids sur l'analyse de robustesse, la précision prédictive attendue est la manière par laquelle l'approche hybride confirme que le lien de similarité modèle-monde n'est pas le fruit du surajustement.

Une technique populaire pour calculer la précision prédictive attendue est le *Leave-one-*

11. McElreath donne l'exemple du télescope Galilée qui, faute de résolution suffisante, est une source d'incertitude épistémique pour observer les lunes de Saturne pour expliquer pourquoi les bayésiens interprètent le hasard comme une partie intrinsèque de la pratique de modélisation.

out Cross-validation (*loo* ; Vehtari *et al.*, 2017). Loo prend la forme suivante $elpd_{elpd} = \sum_{i=1}^{n} \log \pi_S(\tilde{y}|\tilde{y}_{-i})$, où *elpd* signifie « *expected log pointwise predictive density* », et $\pi_S(\tilde{y}|\tilde{y}_{-i}) = \int \pi_S(\tilde{y}|\theta)\pi_S(\theta|\tilde{y}_i)d\theta$. Avec *loo*, on supprime tour à tour des observations dans un jeu de données, en demandant à chaque fois au modèle de prédire ce point manquant $\pi_(\tilde{y}|\tilde{y}_{-i})$. La prédiction se fait à partir de la probabilité postérieure avec le point manquant, $\theta_S(\theta|\tilde{y}_{-i})$, et la fonction de vraisemblance $\pi_S(\tilde{y}|\theta)$. Cette procédure permet d'estimer les capacités prédictives du modèle sans avoir recours à un nouveau jeu de donnée.

1.3 Résumé

En somme, on a vu deux conceptions de la modélisation qui font différentes prédictions sur la manière dont on peut organiser la grande diversité de modèles observés en science. La conception de Weisberg suggère que seulement trois types de structures de modèles accompagnés de leurs interprétations sont suffisants pour expliquer l'ensemble des pratiques de modélisation. Une conception alternative est de reconnaître l'émergence d'une pratique hybride de modélisation comme un phénomène nouveau et distinct, qui s'accompagne d'une interprétation dans laquelle la frontière entre le théorique et l'empirique est plus poreuse. Dans la pratique hybride, la cible est le processus observationnel, lequel est interprété comme nécessaire dans des contextes dits empiriquement difficiles. De plus, la pratique hybride met l'accent sur la résolution du problème inverse, car elle cherche à construire des modèles théoriques capables de produire des prédictions de qualité. L'une des conséquences de la complexité additionnelle des MPPs est qu'ils sont davantage évalués sur leurs capacités d'apprentissage que les modèles théoriques traditionnels.

CHAPITRE II

MÉTHODOLOGIE

Dans le dernier chapitre, on a d'abord abordé la manière dont Weisberg résume les types de modélisation, puis on a argumenté pour interpréter la pratique hybride en tant que pratique distincte. Dans ce deuxième chapitre et les suivants, on confronte ces arguments aux données en faisant une cartographie des pratiques de modélisation en écologie. Comme mentionné dans l'introduction, l'objectif de la cartographie est de situer les différentes pratiques de modélisation les unes par rapport aux autres dans différentes revues scientifiques et à travers le temps. Il s'agit donc d'établir dans quelle mesure la perspective de Weisberg se généralise à un ensemble de données non observées.

L'une des difficultés avec cette cartographie sera de caractériser les éléments qualitatifs du chapitre précédent à l'aide de méthodes quantitatives. Ce passage de la philosophie aux humanités numériques repose sur deux suppositions ; (i) la manière dont les scientifiques discutent de leurs modèles est informative des pratiques de modélisation, et (ii) les différentes significations des mots proviennent de leurs contextes (appelé l'hypothèse distributionnelle ; Harris, 1954 ; Firth, 1957). Ainsi, pris ensemble, on obtient que *l'étude des différentes pratiques de modélisation, et l'évolution de celles-ci, peut passer par l'analyse des contextes dans lesquels les discours sur les modèles prennent place.*

Dans un premier temps, on introduit le corpus en écologie et la manière dont il a été nettoyé pour des fins d'analyses (Sec. 2.1). En plus du corpus principal, on introduit

dans cette section un plus petit corpus de validation composé de textes prototypiques liés aux différentes pratiques de modélisation, utile pour esquisser les contours du territoire d'intérêt. On prend ensuite le temps d'explorer le corpus de validation à l'aide de différentes techniques en traitement naturel du langage (Sec. 2.2). Cette étape d'exploration est nécessaire, car elle permet de préparer le terrain pour valider les résultats des deux techniques dites non supervisées utilisées dans ce mémoire, lesquelles sont présentées dans la section suivante (Sec. 2.3). On conclut sur la manière dont ces techniques illustrent les deux suppositions ci-haut et nous permettent ainsi de cartographier les pratiques de modélisation. [1]

2.1 Le territoire à cartographier

2.1.1 Choix et contenu du corpus

On a acquis le jeu de données par l'interface de programmation d'application de JSTOR Data For Research (DfR), un service mis à la disposition des chercheur.es pour accéder à un corpus d'intérêt. Suite à une enquête préliminaire pour déterminer les revues pertinentes pour notre question de recherche, on a retenu quatre revues en écologie, à savoir *Journal of Animal Ecology, Ecology, Journal of Ecology*, et *Ecological Monographs*.

Le domaine de l'écologie est privilégié en raison de sa riche tradition de modélisation, notamment avec Levins dont nous avons évoqué son importance dans l'introduction. On a choisi les quatre revues ci-dessus, car ce sont des revues phares de l'écologie dans lesquelles il y a un effort de synthèse. Inspirés par des pionniers de l'écologie tels que MacArthur, Hutchinson, May, Holling et bien d'autres, les auteur.es de ces revues

---

1. En plus du document principal, il y a une application en ligne qui permet d'explorer de manière interactive les données du corpus : `http://shiny.initiativesnumeriques.org/ecology-corpus-explorer/`. À différents endroits dans le texte, on propose de visiter cette application pour accéder à des visualisations d'intérêts. On conçoit cette application comme une expérience en soi qui vise à promouvoir différents médiums pour mettre de l'avant les bénéfices des humanités numériques.

utilisent les différents types de modèles discutés dans le premier chapitre pour articuler leurs idées. De plus, on y retrouve plusieurs articles fondateurs qui ont jeté les bases des pratiques de modélisation contemporaines. Ces revues ont également l'avantage d'être publiées depuis les années 1960. On commence notre analyse à cette date notamment parce que plusieurs programmes de recherche contemporains ont été mis en place à l'époque (p.ex. biogéographie des îles, écologie des systèmes), mais aussi pour des raisons pratiques liées au nettoyage du corpus. [2]

2.1.2 Nettoyage et prétraitement

Une fois les données récupérées, on nettoie le corpus afin de le mettre dans une forme digeste pour les algorithmes. En premier lieu, on enlève le bruit engendré par l'algorithme de reconnaissance de caractère optique (OCR) utilisé par JSTOR lors de la conversion des fichiers PDFs vers les fichiers textes. Le bruit auquel on réfère est les informations de bas de page du PDF dans le texte, les barres obliques inverses devant des guillemets, des mots séparés par des tirets, les points d'interrogation en guise de caractères spéciaux, etc (voir l'extrait dans l'appendice). Ce nettoyage inclut aussi les *boilerplate words*, ce qui signifie d'enlever les entêtes, les mots-clés, et les résumés qui se trouvent au début des textes, ainsi que les bibliographies et remerciements à la fin du texte.

Étant donné le manque de standardisation à travers les époques et les revues (chaque époque a ses propres entêtes, habitudes de remerciement, problèmes d'encodage, etc), l'étape de nettoyage prend beaucoup de temps et est, en général, reconnue comme laborieuse. C'est l'une des raisons pour laquelle on préfère commencer l'analyse du corpus à partir des années 1960, malgré le fait que celui-ci recule plus loin dans le temps. Le tableau suivant résume le corpus avant le nettoyage :

---

2. Le temps nécessaire à l'étape de nettoyage est la raison pour laquelle il est difficile d'ajouter de nouvelles revues à notre analyse en cours de route.

Tableau 2.1: Résumé du nombre de mots uniques (numérateur) sur le nombre d'occurrences dans le corpus *avant* le nettoyage (dénominateur).

Revue	1960	1970	1980	1990	2000	2010	Total
Eco. Mono.	94K/1.37M	77K/1.33M	73K/1.30M	82K/1.88M	94K/2.34M	70K/1.55M	280K/9.76M
Ecology	155K/3.33M	152K/4.55M	166K/6.54M	201K/9.21M	231K/12.12M	176K/7.93M	605K/43.68M
Jn of An. Eco.	58K/1.33M	69K/1.80M	67K/2.21M	92K/3.00M	117K/4.28M	78K/2.18M	265K/14.79M
Jn of Ecology	68K/1.53M	78K/2.04M	78K/2.26M	115K/2.88M	146K/4.28M	95K/2.43M	330K/15.43M
Total	258K/7.57M	248K/9.72M	252K/12.31M	321K/16.96M	380K/23.02M	267K/14.09M	1M/83.66M

En plus du nettoyage, il est courant dans l'analyse du langage naturel de «prétraiter» le corpus. Cette étape de prétraitement consiste notamment à enlever les symboles de ponctuation et les nombres, et de mettre le texte en minuscule. Afin de réduire au maximum les mots qui sont perçus comme non pertinents pour l'analyse sémantique ; il est également coutume d'enlever une liste de mots vides du texte. [3]

Une fois prétraité, on *tokenise* le texte de manière à ne garder que les mots, [4] puis on *lemmatise* afin de ramener les mots à une forme réduite, celle du lexème. La lemmatisation est surtout utile avec les plus petits corpus, notamment parce qu'elle évite à l'algorithme d'apprendre les différentes inflexions des mots (Jurafsky et Martin, 2019). Dans notre cas, vu que l'on divise notre corpus par tranche de temps pour effectuer une analyse diachronique sur les plongements lexicaux, cela a pour effet de créer de plus petit corpus pour chacune de nos analyses. Pour des fins de comparaisons, on garde le texte lemmatisé pour la modélisation thématique aussi.

Afin de mettre l'accent sur les pratiques de modélisation, on choisit de travailler avec un sous-ensemble d'articles contenant au moins une occurrence du concept de modèle. Pour ce faire, on sépare d'abord les articles en sections de 500 mots, à savoir environ

3. On considère souvent comme inutiles les mots tels que*the*»,*a*»,*their*»,*by*», etc.

4. La tokenization consiste à séparer un texte en ses éléments constitutifs. Nous travaillerons avec des mots, mais il est aussi possible de travailler avec des n-grammes d'ordre supérieur comme le bi-gramme ou le tri-gramme, etc.

le niveau de la page (l'OCR ne conserve pas les marqueurs de séparation de section). Ensuite, on identifie les sections qui contiennent une occurrence quelconque du lexème de modèle, comme *modeling* ou *models*. Finalement, on reconstruit les articles avec les sections que l'on sait avoir une occurrence de modèles. Pour éviter que certains articles reconstruits soient excessivement plus longs que d'autres, ce qui est problématique pour la modélisation thématique (Airoldi, 2014), on a fixé un seuil de 4000 mots maximum par article. Ce sous-ensemble a l'avantage d'avoir une plus grande proportion du contenu sémantique lié aux modèles, mais introduit potentiellement certains biais. Par exemple, les articles méthodologiques risquent d'être surreprésentés dans l'analyse sémantique. On croit que le grand nombre de données permet quand même d'avoir un portrait représentatif de la production scientifique en lien avec la modélisation. On résume le corpus nettoyé et prétraité avec les deux tableaux suivants :

Tableau 2.2: Résumé du nombre de mots uniques (numérateur) sur le nombre d'occurrences dans le corpus *après* le nettoyage (dénominateur)

Revue	1960	1970	1980	1990	2000	2010	total
Ecological Monographs	7K/31K	12K/81K	14K/112K	24K/350K	38K/631K	36K/539K	77K/2M
Ecology	14K/93K	28K/358K	41K/711K	64K/2M	99K/3M	91K/3M	199K/8M
Journal of Animal Ecology	6K/35K	15K/181K	21K/308K	35K/641K	54K/1M	43K/721K	101K/3M
Journal of Ecology	5K/20K	10K/64K	15K/138K	25K/335K	44K/834K	35K/602K	78K/2M
total	20K/179K	41K/686K	57K/1M	96K/3M	154K/6M	134K/4M	307K/15M

Tableau 2.3: Résumé du nombre de sections (numérateur) sur le nombre d'articles (dénominateur) dans le corpus *après* le nettoyage. On termine avec un total de 14,881 articles, comprenant 94,352 sections.

Revue	1960	1970	1980	1990	2000	2010	total
Ecological Monographs	176/54	356/93	536/120	1485/181	2436/286	1882/194	6871/928
Ecology	774/270	2853/633	5541/1183	12016/1777	21414/2783	16708/2023	59306/8669
Journal of Animal Ecology	228/64	1215/226	1967/417	4110/638	7628/1017	4114/518	19262/2880
Journal of Ecology	104/53	331/153	670/286	1483/502	3642/878	2683/532	8913/2404
total	1282/441	4755/1105	8714/2006	19094/3098	35120/4964	25387/3267	94352/14881

Si on compare le nombre d'occurrences de mots (*tokens*) avant et après le nettoyage,

on termine avec le 1/5 du corpus original (83.66M $\Longrightarrow$ 15M). Ce ratio varie à travers le temps et les revues, allant de 7 fois moins d'occurrences pour la revue *Journal of Ecology* (15.4M $\Longrightarrow$ 2M) à 4.8 fois moins d'occurrences pour la revue *Ecological Monographs* (9.7M $\Longrightarrow$ 2M), et de 47 fois moins d'occurrences pour les années 1960 (7.5M $\Longrightarrow$ 179K) à 3.8 fois moins pour les années 2000 (23M $\Longrightarrow$ 6M). On note que la revue *Ecology* a nettement plus d'occurrence et de documents que les autres, peu importe la décennie. On peut aussi calculer le ratio types/tokens, lequel est une mesure de diversité lexicale de base, et qui passe de 0.11 à 0.03 de 1960 à 2010. Dans l'ensemble, on peut affirmer qu'il y a plus d'articles contenant des occurrences de modèles au fil du temps, et certaines revues discutent plus de modèles que d'autres. [5]

2.2 Un corpus de validation pour mieux s'orienter

En plus du corpus principal, on construit un corpus de petite taille qui contient des textes prototypiques liés aux pratiques de la modélisation mathématique, statistique et hybride. Ce sont des ouvrages de référence que l'on considère comme des données de vérification (*ground truth*), utiles pour comparer avec des résultats d'analyse non supervisée. Pour la modélisation mathématique, on a choisi l'ouvrage de référence *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution* par Otto et Day (2007), pour la modélisation computationnelle on a *Agent-Based and Individual-Based Modeling* par Railsback & Grimm (2012), alors que pour les statistiques, on a *The analysis of biological data* par Whitlock et Schluter (2014). Pour la pratique hybride, on a *Bayesian Models : A Statistical Primer for Ecologists* de Hobbs et Hooten (2015) ainsi qu'un chapitre de McElreath (2020) tiré de son livre *Statistical Rethinking*, intitulé *Generalized Linear Madness*. Pour des fins de comparaison entre les statistiques bayésiennes plus traditionnelles et la pratique hybride, on a aussi le livre de Kruschke

5. D'autres statistiques et visualisations sont accessibles en ligne : `http://shiny.initiativesnumeriques.org/ecology-corpus-explorer/` sous l'onglet *descriptive stats*.

(2018) intitulé *Doing Bayesian Data Analysis*. Afin de mieux s'orienter dans les analyses sémantiques du prochain chapitre, on prend le temps d'explorer le corpus de validation.

On commence par examiner la fréquence des mots dans un document. La fréquence des mots est simplement le nombre d'occurrences d'un mot dans un document sur le nombre total de fois où le mot apparaît dans l'ouvrage. Dans ce cas-ci, on considère chaque chapitre d'un texte protypique comme un document. Ce niveau d'agrégation est celui qui se rapproche le plus du niveau des articles que nous étudierons au prochain chapitre. On obtient les résultats suivants pour le livre d'Otto et Day en modélisation mathématique :

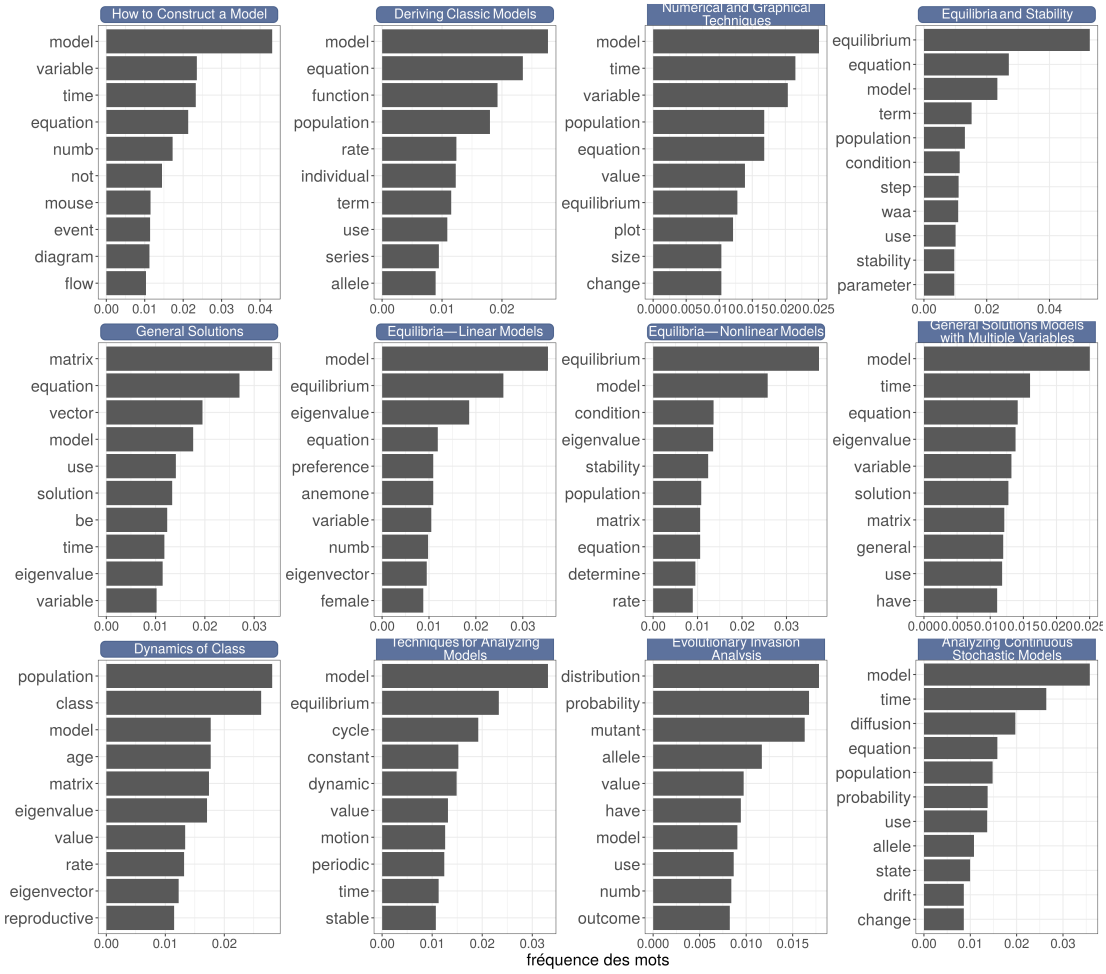


Figure 2.1: Les mots les plus fréquents pour chacun des chapitres du livre de modélisation mathématique en biologie de Otto & Day. Ces termes seront utilisés comme données de vérification pour identifier les thèmes liés à la modélisation mathématique.

Afin d'avoir un point de comparaison, on peut faire de même avec le livre de Whitlock et Schluter sur les statistiques (Figure 2.2), et les textes de Hobbs et Hooten et McElreath (Figure 2.3), lesquels exemplifient l'approche hybride.

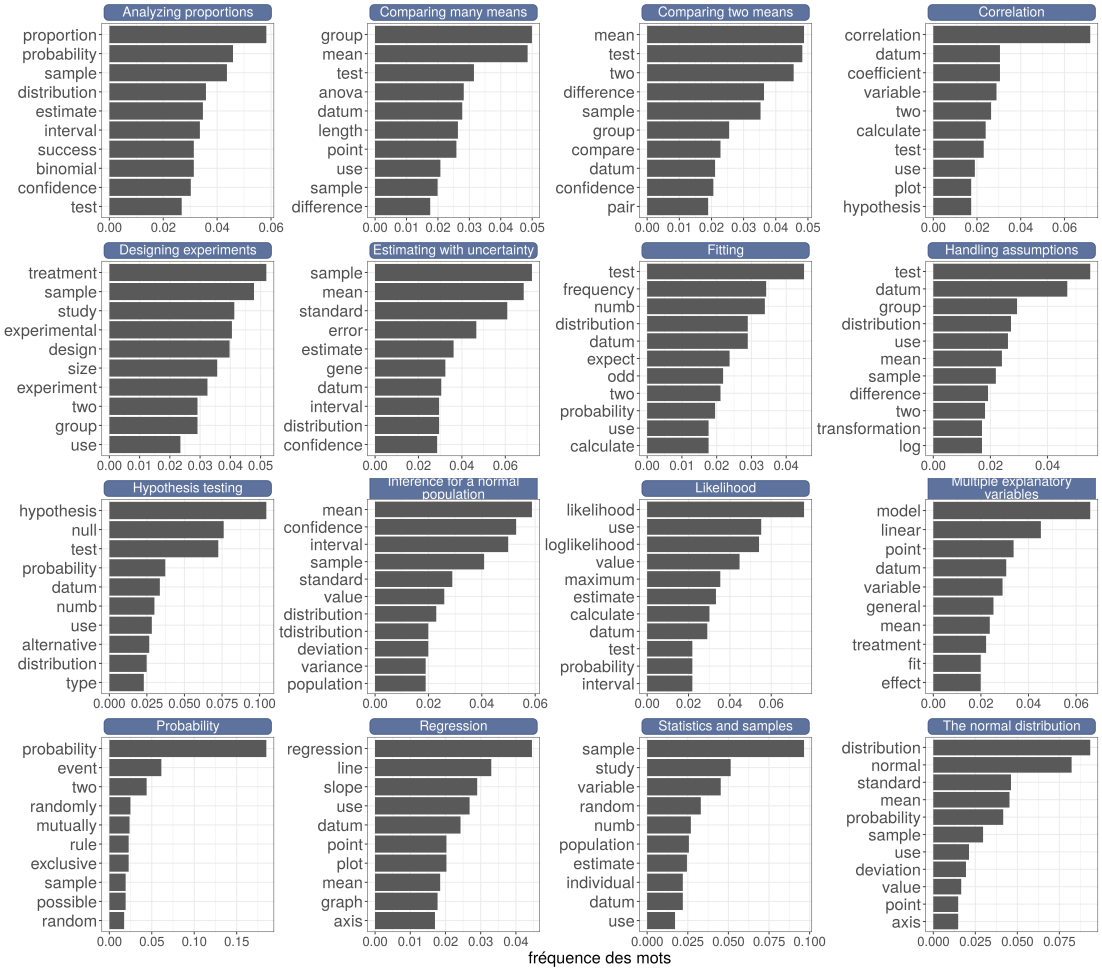


Figure 2.2: Les mots les plus fréquents pour chacun des chapitres du livre de modélisation statistique de Whitlock & Schluter. Tout comme avec la modélisation mathématique, ces termes seront la base pour identifier les thèmes de modélisation statistique.

On observe d'abord dans la Figure 2.1 que les chapitres d'Otto et Day discutent fréquemment des *equilibrium state*, ce qui est courant pour l'analyse dynamique d'un système complexe à l'aide d'équations différentielles. En général, on note aussi plusieurs termes en lien avec la temporalité, à savoir *rate, flow, cycle, stability*, et la dynamique des populations. Il y a deux chapitres qui utilisent grandement des notions matricielles de *eigenvalue* et *eigenvector*, et vers la fin, deux chapitres qui utilisent des notions de probabilité. On remarque également la présence de termes en lien avec l'écologie ici et là,

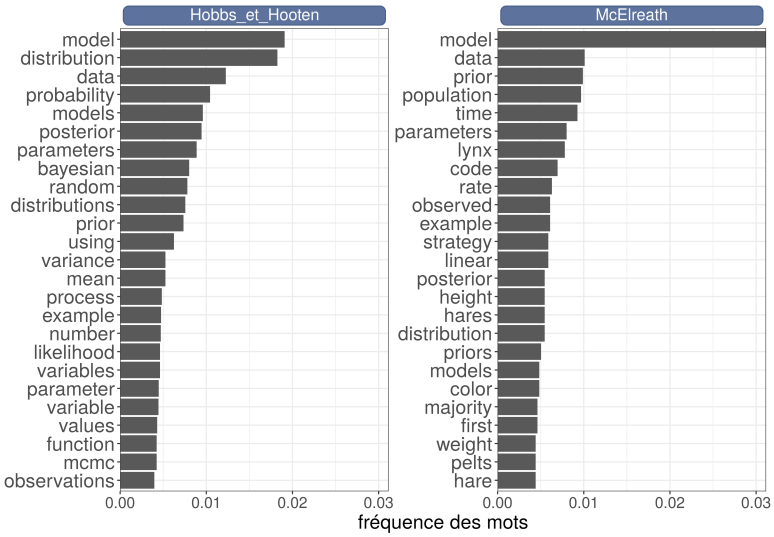


Figure 2.3: Les mots les plus fréquents de l'ouvrage de Hobbs & Hooten et d'un chapitre du livre de McEleath, lesquels on considère comme représentatif de la pratique hybride. La pratique hybride ne disposant pas d'un ouvrage de référence bien établi, des textes de sources différentes sont utilisés pour représenter cette pratique.

à savoir *anemone*, *allele*, *reproductive*, ou *mouse*.

Si on dirige notre attention sur le texte de Whitlock et Schluter (Figure 2.2), on remarque sans surprise l'importance des notions de *datum*, *mean*, et de *sampling*. On peut facilement distinguer les notions en lien au domaine l'expérimental et les tests d'hypothèses (p.ex., *treatment, design, group,anova, ...*), essentiellement la première partie du livre, et les notions en lien avec la modélisation statistique plus généralement, à savoir les concepts tels que *regression, likelihood, estimation*, et *linear model*. En somme, l'ouvrage de statistique se caractérise notamment par des termes en lien avec les données regroupées et la quantification des erreurs d'observations.

Finalement, dans les textes de Hobbs et Hooten et McElreath (Figure 2.3), on note que ceux-ci possèdent un mélange des termes des deux premières visualisations. Par exemple, on retrouve les termes *process, rate, lynx/hare, dynamics, density*, et *state*, lesquels sont des termes importants notamment dans le chapitres sur les modèles classiques et celui de la dynamique des populations de Otto and Day, ainsi que les termes tels que *data,*

distribution, variance, mean, et *linear*, lesquels sont des termes surtout présents en statistiques. Ce type d'amalgame est caractérisé par des configurations de mots, qui combinent modélisation mathématique et statistique. C'est ce que nous chercherons à faire apparaître dans le corpus principal avec les différentes analyses sémantiques.

Bien que la fréquence des mots soit utile pour explorer un corpus, celle-ci reste imparfaite à bien des égards. Par exemple, les mots les plus souvent utilisés sont importants, mais ils ne sont pas nécessairement plus *représentatifs* d'un document pour autant. De plus, c'est une représentation qui reste très grossière dans le sens où les relations sémantiques entre les différents mots ne sont pas prises en compte. Une façon alternative d'explorer le corpus est d'examiner le *tf-idf*, qui est une mesure de représentativé sémantique (Jurafsky et Martin, 2019). Le tf-idf prend la forme suivante :

$$w_{\mathrm{t,d}} = \mathrm{tf}_{\mathrm{t,d}} \text{ x } \mathrm{idf}_{\mathrm{t}}$$

où la pondération w d'un mot t dans un document d est la combinaison de la fréquence et du log de la fréquence inverse, à savoir $\ln(n_{\mathrm{d}}/n_{\mathrm{d \text{ avec une occurrence de t}}})$. Autrement dit, un mot est représentatif s'il est à la fois fréquent dans un document et peu utilisé dans les autres documents. Si on considère chaque ouvrage de référence comme un document, les termes les plus représentatifs sont les suivants :

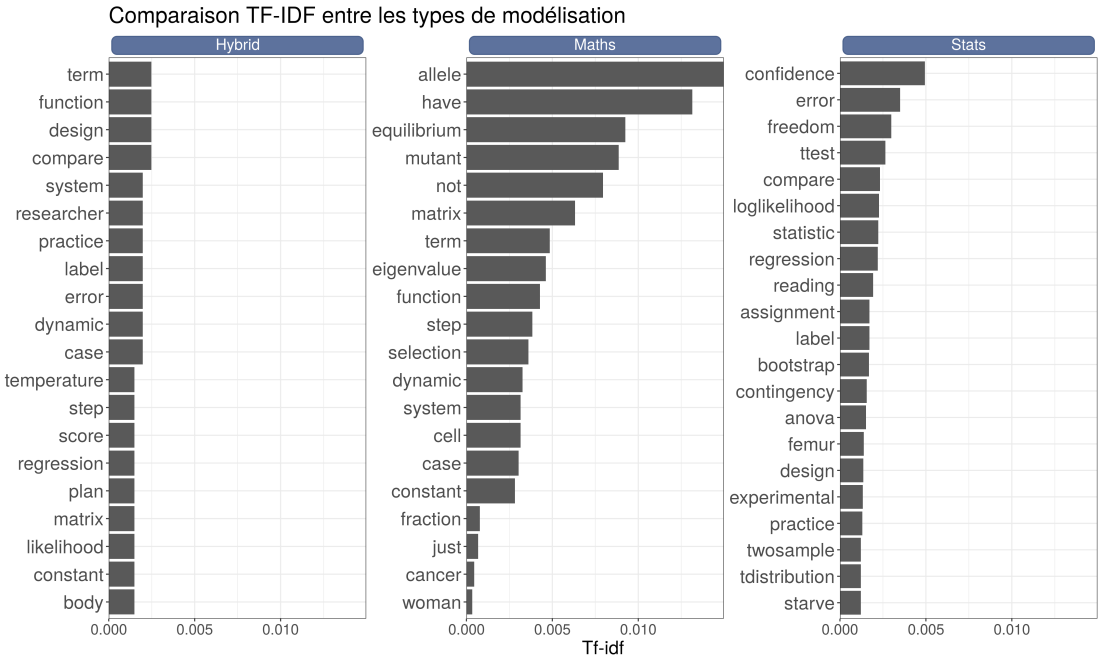


Figure 2.4: Un exemple du tf-idf qui compare les documents étiquetés comme hybrides, mathématiques ou statistiques. Dans cette figure, on utilise le tf-idf pour mettre en évidence les termes les plus représentatifs des différentes pratiques de modélisation, par rapport aux autres types.

Dans l'ensemble, on note que l'on retrouve en partie les termes propres à chaque pratique établis précédemment, mais de manière différente. On peut affirmer que la modélisation statistique se distingue des autres pratiques par la dimension expérimentale, les modèles de régression, et les tests statistiques, alors que la modélisation mathématique se distingue par l'étude des dynamiques des systèmes et la présence de discussion sur les connaissances de domaine. On remarque que les termes représentatifs de la modélisation hybride combinent des termes sur la dynamique, les connaissances de domaine, et les modèles de régression, mais les valeurs du tf-idf sont moins prononcées que dans les deux autres pratiques. Ce résultat reflète la difficulté de démontrer l'existence de la pratique hybride, qui, on suggère, est occultée par les deux autres pratiques.

Enfin, il est possible de mettre en évidence la similarité relative des différentes pratiques avec la mesure du *cosine*. Avec cette mesure de similarité, chaque document est repré-

senté par un vecteur de fréquence de mots. La similarité cosine entre deux documents consiste à prendre le produit scalaire de leurs vecteurs divisé par le produit de leurs normes. Les coordonnées des documents sont données par la matrice document-termes du petit corpus, où chaque document est une rangée et chaque mot est une colonne. Si on interprète les liens de similarité entre les documents comme des arêtes, et les documents comme des noeuds, on obtient un réseau de similarité sémantique :

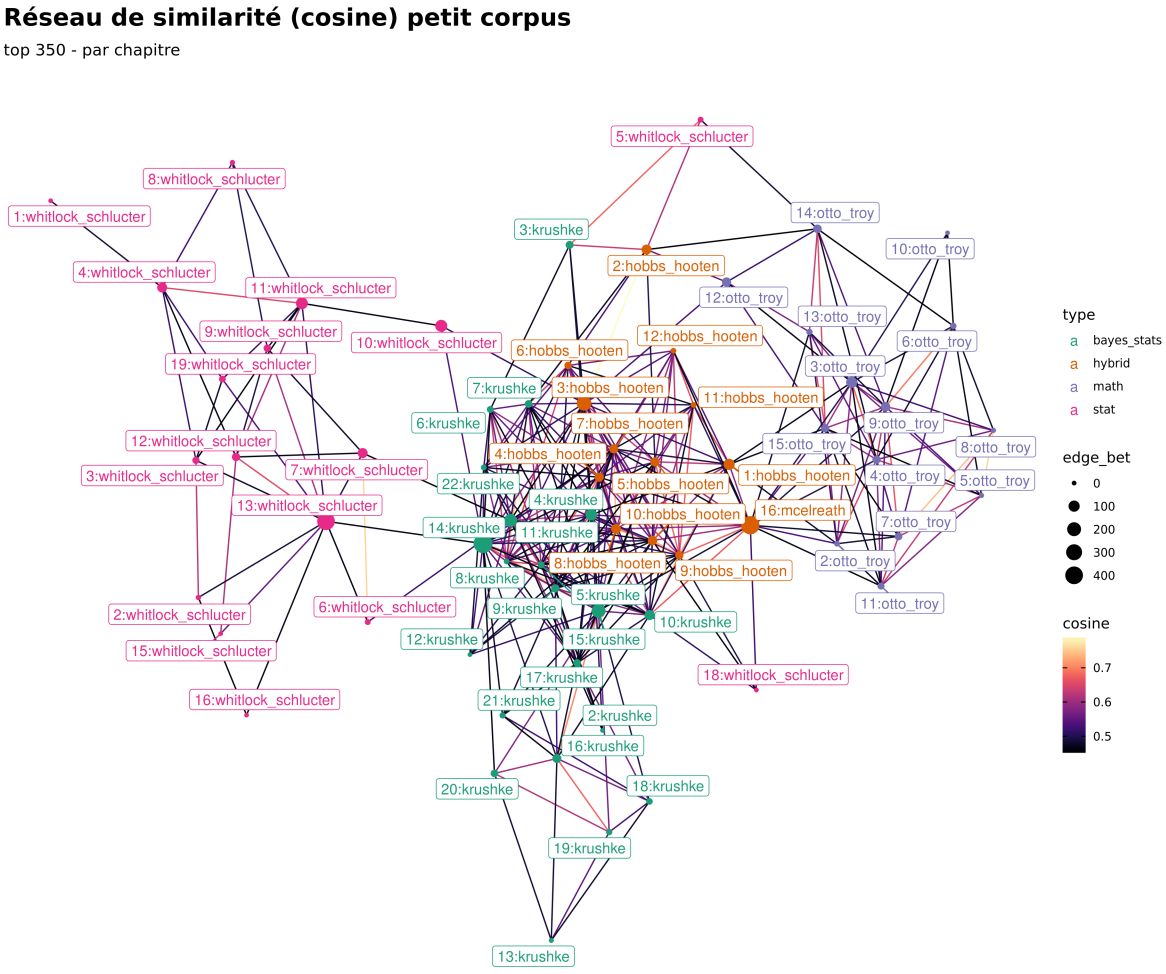


Figure 2.5: Réseau de similarité entre des documents prototypiques des différentes pratiques de modélisation.

Dans la Figure 2.5, on constate l'éloignement sémantique entre la pratique de modélisation mathématique et celle de la modélisation statistique. On note certaines exceptions, tel que le chapitre 5 du livre de Whitlock et Schluter qui porte sur les probabilités. On note aussi que les œuvres choisies comme représentatives de la statistique bayésienne et de la pratique hybride se positionnent confortablement entre les deux pratiques, avec la pratique hybride qui tend à utiliser davantage le langage de la modélisation mathématique. Parmi les différents chapitres de la pratique hybride, le chapitre 16 de McElreath en particulier joue le rôle d'intermédiaire entre les statistiques bayésiennes et la modélisation mathématique. L'intermédiarité défini ici comme le nombre de fois où d'autres dyades de nœuds doivent passer par ce nœud pour emprunter le chemin le plus court.

Bien que l'analyse exploratoire suggère des pistes de recherche intéressantes, la prudence est de mise avec les techniques à base de comptage. Par exemple, on s'imagine que certains termes comme *distribution* ou *model* sont hautement polysémiques. Malgré leur définition formelle, ceux-ci sont utilisés de manières différentes dans différents contextes. Il a été démontré que le sens des termes polysémiques est sujet à une plus grande évolution (Hamilton *et al.*, 2018), et donc il n'est pas surprenant de le voir apparaître plus fréquemment lorsque l'on évalue leur représentativité avec des techniques à base de compte. Afin de comprendre les subtilités des relations sémantiques du corpus, il est nécessaire de les modéliser. C'est pourquoi on distingue ces outils d'exploration des modèles sémantiques que l'on utilise pour la cartographie.

2.3 Modèles d'analyses sémantiques

On fait appel à deux familles de méthodes computationnelles pour étudier la place relative et l'évolution des différentes pratiques de modélisation. La première famille de méthodes est associée à l'allocation de Dirichlet latente (LDA), laquelle est une technique non supervisée qui permet de regrouper des termes qui appartiennent à une même structure latente. Étant donnée la nature non supervisée de cette approche, elle

est utilisée de manière à comparer notre interprétation des pratiques de modélisation à une classification dite *bottom-up*. La seconde technique utilisée est la représentation du plongement lexical (*word embedding*), et en particulier le modèle *word2vec*, qui est grandement utile pour représenter les mots en fonction des relations de cooccurrences. Le plongement lexical se prête volontiers à la représentation de réseau de termes associés au concept de « modèle ». En examinant l'évolution de ce graphe de plongements lexicaux, on peut vérifier si la dynamique d'utilisation des modèles correspond aux modèles présentés dans le premier chapitre.

2.3.1 La modélisation thématique

On introduit d'abord la LDA (Blei, Ng, and Jordan 2003 ; Blei 2012) et le CTM (Blei et Lafferty, 2007), puis la variante que l'on utilisera, à savoir la modélisation par thème structuré (STM). On discute des enjeux d'évaluation, de sélection, et d'interprétation en analyse thématique.

2.3.2 LDA, CTM, STM, alouette

La LDA est un modèle de regroupement qui est grandement utilisé en humanités numériques pour déterminer, de manière non supervisée, les sujets latents présents dans un corpus. La LDA se caractérise par les deux suppositions suivantes ; (i) à chaque document est associé une distribution de probabilité des thèmes, et (ii) à chaque thème est associée une distribution de probabilité de mots. Dans notre cas, cela pourrait être que le document 2343 est un mélange du thème hôtes et parasites à 63%, celui de régression normale à 23%, et finalement du thème échantillonnage à 14%. Un enjeu méthodologique récurrent avec la LDA est la nécessité d'étiqueter les thèmes, i.e., mettre une étiquette qui résume ce que la structure latente qu'un thème est censé représenter. Ces étiquettes ne sont pas déterminées par l'algorithme, mais par l'agent selon son interprétation des mots les plus fréquents des thèmes.

La LDA suppose que la présence ou l'absence d'un thème dans un document est indépendante des autres thèmes qui y figurent, ce qui n'est pas le cas. Par exemple, lorsqu'un document mentionne l'utilisation de tests d'hypothèse, il est probable que l'on retrouve aussi un thème en lien avec les procédures expérimentales. Pour capturer cette dépendance entre les thèmes, Blei et Lafferty (2007) ont développé le *correlated topic model* (CTM). Pour ce faire, le CTM remplace la distribution de Dirichlet dans la LDA avec une distribution logistique-normale (Aitchison et Shen, 1980). Tout comme avec la loi normale multidimensionnelle, la logistique-normale représente la corrélation entre les thèmes à l'aide d'une matrice de covariance, laquelle est par la suite projetée sur un *simplex* afin d'obtenir la prévalence des thèmes sous forme de proportion (voir *appendice B* pour les détails techniques). À l'aide de la vraisemblance retenue (*held-out likelihood*), une méthode d'évaluation des modèles qu'on explique dans la prochaine section, Blei et Lafferty ont démontré que l'inclusion de la corrélation entre les thèmes améliore les capacités prédictives du CTM par rapport à la LDA (Blei et Lafferty, 2007). En particulier, le CTM requiert moins de données avant de pouvoir mieux se généraliser sur la partie non observée.

La STM généralise le CTM en permettant l'inclusion de la corrélation entre les thèmes pour modéliser la prévalence des thèmes, mais aussi d'autres métadonnées que les scientifiques pensent pertinentes (Roberts *et al.*, 2016; Roberts *et al.*, 2019). Concrètement, la STM introduit un modèle de régression linéaire dans la CTM afin de pouvoir tester l'influence de différentes variables sur la prévalence des thèmes. Tout comme la CTM, les auteur.es de la STM démontrent que l'addition d'information dans le modèle permet d'améliorer la capacité prédictive du modèle, à condition que les variables soient informatives sur la prévalence des thèmes. Dans notre cas, on ajoute la date de publication des articles et la provenance des revues. On s'attend à ce que les documents de la même revue et de la même époque aient tendance à se ressembler davantage que ceux qui ne le sont pas.

En plus d'améliorer les capacités prédictives du modèle, la structure de dépendance entre les thèmes se représente aisément sous forme de graphe où les nœuds sont des thèmes et les arêtes sont les relations de corrélation. Ce « réseau de thèmes » permet de visualiser, ou cartographier selon notre analogie, le paysage global d'une revue et met en évidence les relations entre des thèmes qui sont invisibles lorsqu'on se limite au contenu de thèmes individuels (Blei et Lafferty, 2007). De plus, on examine les plus fortes relations pour identifier les regroupements sémantiques présents dans le réseau de thèmes, ce que l'on réfère comme communautés sémantiques. Les communautés sémantiques constituent l'une des structures qui se rapprochent le plus des pratiques de modélisation définies dans le premier chapitre du mémoire.

Comment savons-nous si la STM donne de bons résultats ?

L'une des principales difficultés avec les modèles non supervisés concerne leur sélection et leur validation. Pour faire écho aux questions du premier chapitre, y a-t-il un meilleur modèle ? Est-ce qu'il y a un bon nombre de thèmes à choisir ? Est-ce ce que le modèle qui démontre la meilleure capacité prédictive devrait être nécessairement favorisé, au détriment du jugement humain ? On aborde ces questions tour à tour.

La question du choix du bon nombre de thèmes est délicate, car il n'y a pas une meilleure réponse (Grimmer et Stewart, 2013). Lorsqu'on compare un même corpus à 25 et 150 thèmes, on retrouve plusieurs thèmes à $K = 150$ qui sont des décompositions plus fines des thèmes à $K = 25$. Pour choisir le bon nombre de thèmes, il est donc important de considérer le niveau d'analyse qui est propre à nos intérêts (Malaterre, Chartier, and Pulizzotto 2019). Cela dit, il existe néanmoins un bon nombre de thèmes à partir desquels les capacités prédictives du modèle sont optimisées.

Pour analyser la capacité prédictive des modèles thématiques, on peut faire appel à la vraisemblance retenue (Wallach *et al.*, 2009). La vraisemblance retenue a pour but

d'estimer la précision prédictive attendue du modèle qui, rappelons-le, vise à évaluer la capacité d'un modèle à généraliser son apprentissage sur des données non observées. Pour ce faire, on supprime la moitié des mots dans 10% des articles du corpus, et on demande à l'algorithme de prédire les mots manquants, compte tenu du texte observé dans la première partie du document. Ce faisant, si on spécifie un trop grand nombre de thèmes pour les données en notre possession, chaque thème devient extrêmement granulaire (au point où il serait calqué sur le texte). Dans ce cas, le modèle ne parvient pas à apprendre la structure latente du document et échoue à prédire correctement les mots manquants. L'intuition est similaire pour un trop petit nombre de thèmes.

Bien que la capacité prédictive d'un thème soit importante pour évaluer un modèle, il est toujours recommandé de trouver un équilibre entre l'interprétabilité et la capacité prédictive des sujets (Airoldi et Bischof, 2016). Ainsi, il faut être à la fois ouvert à la possibilité que les algorithmes découvrent des thèmes surprenants tout en faisant appel à l'expertise de domaine quand il vient le temps de juger de la plausibilité et de la qualité des thèmes. Cet exercice d'équilibre nous amène sur le terrain de l'interprétation des modèles, que nous considérons comme faisant partie de l'étape de validation et de sélection des modèles.

Pour interpréter les modèles, on utilise les mesures de la cohérence sémantique et l'exclusivité des thèmes (Bischof et Airoldi, 2012). Comme son nom l'indique, la cohérence sémantique cherche à quantifier notre intuition que certains ensembles de mots semblent plus cohérents que d'autres (Mimno *et al.*, 2011). Contrairement au score de vraisemblance retenue, cette mesure a l'avantage d'avoir une dimension empirique. En effet, plusieurs expériences ont été menées pour quantifier la similarité entre le jugement expert et la cohérence sémantique. Par exemple, plusieurs experts de l'Institut National de la Santé américain NIH) ont été réunis afin d'évaluer la qualité des thèmes d'une LDA entraînée sur 300,000 demandes de subventions et résumés d'articles (Mimno *et al.*, 2011). Les experts devaient catégoriser les thèmes dans différentes catégories et, lors-

qu'un thème est jugé de «mauvaise qualité», ceux-ci devaient diagnostiquer les problèmes du thème. Il s'avère que la cohérence sémantique'est fortement corrélée avec la qualité perçue des thèmes par les experts.

Un enjeu avec la cohérence sémantique est qu'elle est facile à réaliser lorsqu'un sujet est dominé par quelques mots (Roberts *et al.*, 2014). Une solution proposée par Roberts et coll. est de prendre en compte l'exclusivité des mots dans un thème. Pour ce faire, on utilise l'indice de Fréquence-Exclusivité, ou FREX (Bischof and Airoldi 2012 ; Airoldi and Bischof 2016), qui est une moyenne harmonique pondérée dans laquelle le rang du mot est une combinaison de sa fréquence et de son exclusivité. [6] En raison de la prépondérance du terme *model* dans la modélisation thématique, on utilise fréquemment le score FREX pour l'interprétation des thèmes.

Il reste que la modélisation thématique a certaines limites en ce qui a trait à l'étude de l'évolution des modèles. Bien que l'on puisse mesurer l'évolution des thèmes, la STM ne modélise pas explicitement l'évolution sémantique des mots. Par exemple, on s'imagine que le terme *model* se trouve à la fois dans un thème lié à la dynamique des populations et à la régression, et que ce dernier a gagné en popularité aux dépens du premier. Bien que ce changement suggère que le sens du terme *model* s'est probablement rapproché de celui de *regression*, ce changement de sens n'est pas directement modélisé par la STM. Ce changement de proportion de thèmes est une preuve indirecte de l'évolution sémantique du terme modèle. Pour revenir à la question de recherche, dans quelle mesure le terme *model* est-il davantage statistique que mathématique ? Ou, comme on le suggère, dans quelle mesure est-ce que le sens de *model* s'est rapproché sémantiquement d'un mélange hybride de termes mathématiques et statistiques. Le plongement lexical permet d'avoir un grain d'analyse plus fin et est utilisé comme un outil complémentaire à la modélisation thématique.

---

6. Les détails mathématiques du calcul de la cohérence sémantique et de l'indice FREX se trouvent dans l'appendice B.

Comment savons-nous si les thèmes sont théoriques ou empiriques ?

Afin de classifier les thèmes en fonction des documents du corpus de validation, on calcule la similarité moyenne des thèmes avec les chapitres des ouvrages de références d'intérêt. Pour ce faire, on identifie d'abord les 20 termes les plus représentatifs des thèmes, en fonction du score FREX, ainsi que les 20 termes les plus fréquents pour chacun des chapitres des ouvrages de référence choisis. Afin de distinguer les différents types de modélisation, on compare les ouvrages d'Otto & Day, Railsback & Grimm, et de Whitlock & Schluter. Ensuite, on crée une matrice terme-documents dans laquelle les rangées sont les documents (les chapitres des ouvrages de référence et les 150 thèmes trouvés), les colonnes sont les mots, et les valeurs des cellules sont la fréquence des mots pour les ouvrages de référence et les valeurs bêta des thèmes. À l'aide de cette matrice terme-document, on calcule la similarité des chapitres avec les thèmes à l'aide de la similarité cosinus, puis on prend la moyenne de la similarité des thèmes pour chaque chapitre. On obtient ainsi une matrice qui contient la similarité moyenne des thèmes avec les différents documents du corpus de validation. Finalement, on classifie les thèmes selon les plus fortes relations de similarité.

2.3.3 Réseau des plongements lexicaux

Le plongement lexical est une manière de représenter les données textuelles à l'aide de vecteurs compacts, lesquels encodent les relations sémantiques à partir du contexte des mots (Jurafsky et Martin, 2019). Initialement développé comme représentation distribuée pour les réseaux de neurones (Bengio *et al.*, 2003), le plongement lexical permet de représenter la dimensionnalité élevée des données textuelles dans un espace de basse dimension. Un avantage de cet espace en basse dimension est que la distance entre les mots devient significative et interprétable. C'est pourquoi il est commun d'utiliser des tâches d'« algèbre sémantique » pour valider l'apprentissage de l'espace vectoriel. Par

exemple, si on prend le vecteur du terme *theory* dans le plongement lexical entrainé sur notre corpus, on soustrait celui de *prediction*, et on ajoute celui de *hypothesis*, on obtient les termes suivants :

Term	Similarity	Rank
idea	0.86	1
concept	0.81	2
principe	0.80	3

On apprend que des théories associés à des hypothèses mais sans prédictions sont similaires simplement à des idées, des concepts ou des principes ! On utilise cet espace de basse dimension pour étudier les différents contextes d'utilisations des termes d'intérêts comme *model* ou *bayesian.*[7]

Afin d'explorer les différents usages des modèles, on développe une méthode expérimentale originale inspirée de l'étude des structures de parenté en anthropologie (Bonvillain, 2010). On représente les plongements lexicaux sous forme de graphe où les noeuds sont les mots, et les arêtes sont leurs relations de similarité. Ce graphe de plongements lexicaux est similaire à un réseau de liens de parenté en anthropologie, où on s'intéresse à un individu en particulier, à savoir *model* (appelé « *ego* » en anthropologie). Tout comme en anthropologie, on s'intéresse aux relations de premier et de second ordre pour mieux comprendre la place d'un individu dans le réseau d'interaction. Il est courant de ne pas représenter explicitement ego dans la visualisation, car l'information est redondante, i.e., on sait qu'il s'agit déjà du graphe de ses relations (Crossley *et al.*, 2015). Cette manière de représenter le plongement lexical permet de faire apparaître des regroupements liés aux différentes utilisations du terme *model* dans un contexte donné.

On mentionne la nécessité de choisir un seuil de similarité à partir duquel on considère les termes comme de la parenté dans cette visualisation. On construit le seuil en deux

7. On présente une introduction plus détaillée du plongement lexical dans l'appendice B

temps. D'abord, on choisi les 60 noeuds, ou termes, les plus similaires à ego. Puis, on garde les cinq termes les plus similaires à ces derniers. Cette approche à l'avantage de générer des graphes qui sont modérément connectés et qui permettent l'émergence de regroupements de mots, lesquels se rapprochent des champs sémantiques trouvés par la modélisation thématique. De plus, une fois que l'on a le graphe, on peut utiliser des propriétés des réseau pour faciliter l'interprétation. Pour souligner les termes les plus centraux du graphe, on utilise la mesure de centralité *pagerank* pour mettre de l'avant les termes qui jouent un rôle d'importance dans le graphe. La centralité pagerank est une mesure bien connue qui est utilisée par le moteur de recherche google pour connaître les pages web les plus importantes du graphe de l'internet.

Afin d'étudier l'évolution sémantique, on découpe le corpus par tranches de temps et on examine comment le sens des mots change d'une fenêtre temporelle à l'autre. [8] En particulier, on veut connaître l'évolution relative des mots d'intérêts, c'est-à-dire comment des mots associés aux différentes pratiques de modélisation se rapprochent ou s'éloignent à travers le temps. Par exemple, on sait que le terme *prior* à grandement changé depuis les années 1960. Bien qu'au début le terme ait le sens commun de *previously*, il a acquis depuis les années 1990 le sens bayésien qu'on lui connait :

---

8. Cette approche est inspirée des travaux de Hamilton et coll. (Hamilton *et al.*, 2018)

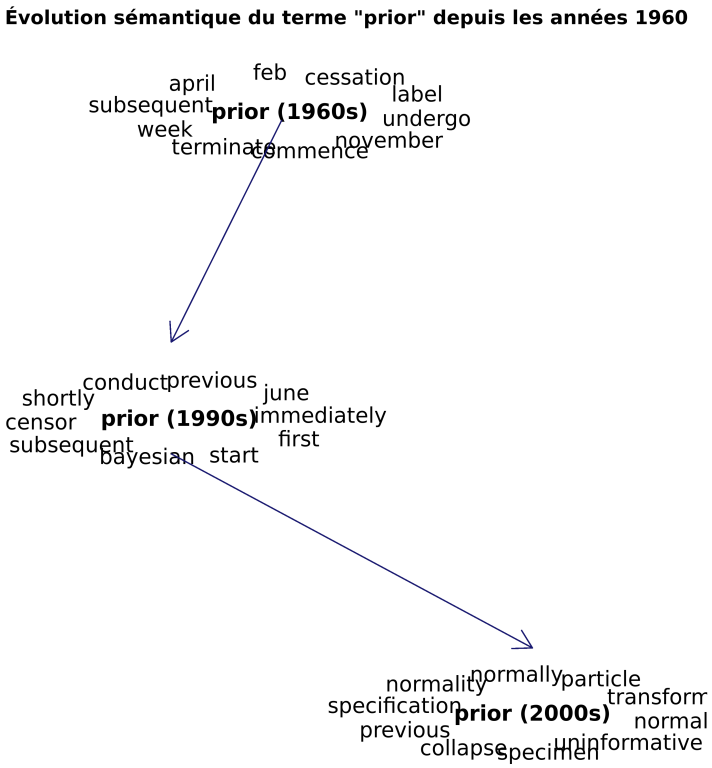


Figure 2.6: Exemple du plongement lexical comme outil diachronique. On voit comment le terme prior a évolué depuis les années 1960 en examinant l'évolution des relations de proximité sémantiques.

Comme pour la modélisation thématique, la validité de cette approche vient de sa capacité à reproduire le jugement humain dans une tranche de temps (avec des tâches synchroniques), ainsi qu'à détecter des épisodes connus de changement sémantique dans le langage courant à l'aide de jeux de données historiques.

2.3.4 La complémentarité de la modélisation thématique et du plongement lexical

Une manière relativement simple de comprendre la complémentarité entre la modélisation thématique et le plongement lexical est d'examiner leurs manières respectives de

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

(a) La matrice termes-documents avec des valeurs données par la représentation du sac de mots

	data	···	**growth**	**matrix**	**neutral**	**rate**	**species**	···
diversity	2	···	2	1	6	5	73	
model	13	···	23	7	4	9	12	
population	8	···	74	8	3	36	13	
simple	11	···	27	2	0	5	8	
time	5	···	23	7	5	11	27	
theory	6	···	4	1	74	1	18	

Figure 2.7: La matrice de coocurrence : (a) les vecteurs de textes (boîtes rouges) sont ordonnés par le nombre de fois que les apparaissent dans un document, alors que (b) les vecteurs de textes sont ordonnés par la cooccurrence entre les termes (dans ce cas-ci, avec une fenêtre de cinq)

représenter le texte, à savoir la représentation du texte comme un « sac de mots » (*bag of words*) et celle comme matrice de cooccurrences des mots. Si l'on considère la Figure 2.7, le principal avantage de la matrice de mots de documents est que l'on garde une trace du document d'où les mots proviennent. Cela permet notamment de comparer la similarité entre les documents, et d'utiliser la modélisation thématique pour représenter la structure latente des différents textes d'un corpus. La matrice de cooccurrences quant à elle encode directement le contexte des mots. Dans l'exemple ci-dessus, qui est un échantillon de la matrice de cooccurrences du texte de Otto & Day, on peut voir que le terme *population* apparaît fréquemment avec *growth* et *rate* mais pas *neutral* et *data*. En représentant cette matrice sous forme de graphe :

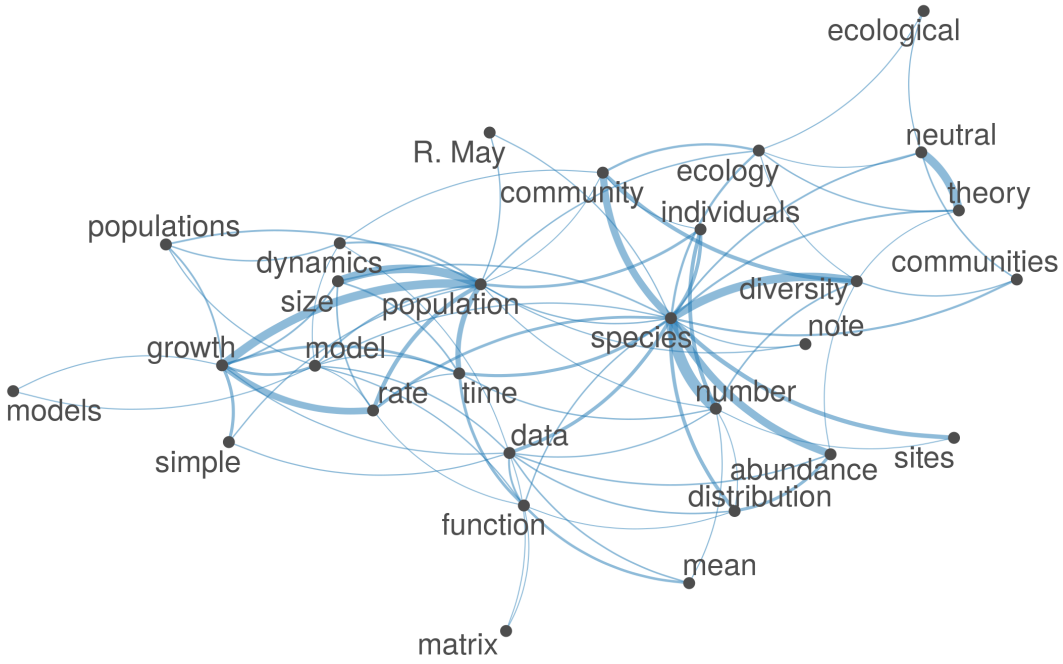


Figure 2.8: Réseau de cooccurrences de l'ouvrage de référence de Otto & Day dans lequel la force du lien est représentée par sa largeur.

on peut facilement examiner les relations de cooccurrence entre les mots dans un corpus. Avec le plongement lexical, il devient possible d'aller au-delà des relations de cooccurrences et de modéliser la proximité sémantique.

Le plongement lexical permet aussi d'étudier l'évolution sémantique des mots en examinant l'évolution des réseaux de plongement lexicaux. On peut donc caractériser la dynamique des différentes pratiques de modélisation en écologie. Cette représentation *bottom-up* permet une analyse temporelle, où chaque graphe est une partition temporelle à un moment donné.

Résumé

Dans ce chapitre, on a introduit le corpus avec lequel nous confronterons les modèles de classification de Weisberg. On a vu comment ce vaste territoire sémantique est influencé

par l'ensemble de nos choix en lien avec le nettoyage, le prétraitement, et la présélection de sections d'intérêts. On a aussi introduit un plus petit corpus, qui, malgré sa taille restreinte, se révèle une partie essentielle de notre cartographie. En effet, celui-ci nous fournit les données de vérification, i.e. les ouvrages de référence représentatifs des différentes pratiques de modélisation. Ainsi, les thèmes mathématiques seront ceux qui sont similaires sémantiquement à l'ouvrage d'Otto & Day (idem pour les autres types de pratiques et leur ouvrage de référence). Afin d'identifier les structures latentes liées aux pratiques de modélisation, on a présenté un modèle d'analyse thématique et un modèle de plongement lexical, à savoir la STM et le modèle word2vec. On cherche à vérifier si les communautés thématiques et l'évolution des relations de similarité sémantique identifiées, respectivement, par la STM et le modèle word2vec sont compatibles avec l'existence d'une pratique hybride.

CHAPITRE III

RÉSULTATS

La conception de Weisberg des modèles scientifiques permet-elle de mieux organiser la grande variété de pratiques de modélisation observées en science ? Observe-t-on une érosion assez importante entre le domaine empirique et théorique pour affirmer, dans certains cas, qu'une nouvelle pratique hybride est un type de modélisation à part entière ? On présente d'abord un aperçu des thèmes en lien avec les pratiques de modélisation (Sec. 3.1), puis on met en relation ces thèmes selon leurs relations de corrélation (Sec. 3.2). Afin de mettre en évidence la robustesse des structures des communautés thématiques, on analyse les patrons d'interactions entre les thèmes à l'aide de trois type d'analyses indépendantes ; une analyse de réseau, un clustering hiérarchique, et une technique de réduction de dimensionnalité (UMAP). On fait valoir que ces structures correspondent aux différents types de modèles, ce qui permet de représenter leur évolution (Sec. 3.3), et donc de vérifier l'existence d'une tendance à l'hybridation des modèles théoriques et empiriques. Afin d'avoir une vision détaillée de certains termes clés des communautés sémantiques, on s'intéresse également à l'évolution des relations de proximité sémantique des termes *model* et *bayesian* par l'entremise des graphes de plongements lexicaux (Sec.3.4).

3.1 Survol des thèmes liés à la modélisation

La combinaison d'une vraisemblance retenue relativement élevée, et d'un compromis entre cohérence sémantique et exclusivité équilibré, suggère que le niveau de granularité à 150 thèmes représente adéquatement le corpus. En effet, ce niveau d'analyse offre une bonne capacité prédictive, ainsi que des thèmes qui sont à la fois sémantiquement cohérents et exclusifs (voir *appendice B* pour plus de détails sur la validation du nombre de thèmes choisis). En plus des méthodes de validation automatiques, une vérification manuelle des thèmes confirme qu'ils sont interprétables et, d'un point de vue écologique, pertinents. Nous nous concentrons sur ce grain d'analyse avec 150 thèmes pour mener à bien notre cartographie des pratiques de modélisation.

Parmi les 150 thèmes analysés, il y en a 46 qui contiennent une occurrence du terme *model* dans leurs 15 mots les plus probables. [1] Au même niveau de granularité, une STM performée sur l'ensemble du corpus identifie quatre thèmes contenant une occurrence du terme *model* dans leurs mots les plus probables. Tous ces thèmes ne sont pas sur les pratiques de modélisation, et certains thèmes discutent des pratiques de modélisation sans les mentionner explicitement. Pour résoudre ce problème, on identifie les thèmes d'intérêts en fonction des termes partagés avec les documents du corpus de validation (Sec. 2.2).

Comme mentionné dans la méthodologie (voir Sec.2.3.2), on établi la proximité entre les thèmes et le corpus de validation sur la base de la similarité cosinus entre les termes représentatifs des thèmes et les chapitres d'ouvrages de référence (en utilisant la valeur bêta et la fréquence des termes, respectivement). Cela dit, il nous reste à établir un seuil de similarité au-delà duquel nous classifions un thème comme mathématique, com-

---

1. Pour explorer les thèmes de manière interactive, rendez vous sur le site `http://shiny.initiativesnumeriques.org/ecology-corpus-explorer/` et cliquer sur l'onglet *Explore Topic Modeling / Words by topic*

putationnel ou statistique. On utilise une stratégie simple, mais efficace, qui consiste à couper dans la queue de la distribution de similarité :

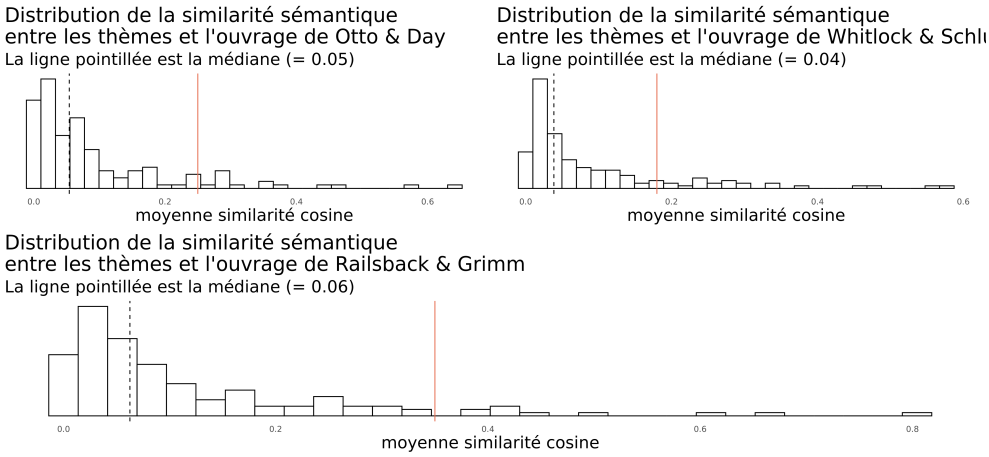


Figure 3.1: Distribution similarité sémantique entre les ouvrages de modélisation mathématique, computationnelle, et statistique et l'ensemble des thèmes. Les lignes orange correspondent aux seuils utilisés.

Concrètement, on trouve qu'un seuil de similarité sémantique moyenne supérieur à 0.25 pour la modélisation mathématique, de 0.35 pour la modélisation computationnelle, et de 0.17 pour la modélisation statistique donne de bons résultats pour identifier les thèmes méthodologiques (voir Figure 3.1).

Cette approche reste heuristique. Bien que certains thèmes soient similaires au corpus de validation, ils ne sont pas pour autant explicitement liés à la modélisation. Par exemple, on trouve que le thème 104 :ASSEMBLAGE [2] est similaire à la fois à la modélisation mathématique et statistique. Une validation manuelle des termes révèle que le thème porte sur le débat autour de la théorie neutraliste de l'évolution (i.e. *null, neutral, assemblage, distribution, random*). Malgré le fait qu'on trouve certains termes proches de la modélisation mathématique et statistique, on juge que la signification des mots dans ce thème

2. Nous utilisons les « small caps » pour indiquer les thèmes

ne correspond pas à celle dans le corpus de validation.[3] Donc, on interprète ce débat comme étant en lien avec la théorie plutôt que méthodologique. De manière analogue, d'autres thèmes comme 78 :DAY, 39 :ENERGY, 26 :EXTINCTION sont à proximité du corpus de validation, mais une validation manuelle révèle que la signification des mots de ces thèmes ne correspondent pas à leur sens méthodologique. Au final, on trouve 35 thèmes (23% des 150 thèmes) qui sont en lien avec les pratiques de modélisation (théorique et empirique), ce qui signifie que la majorité des thèmes identifiés (77%), malgré le nettoyage, sont relatifs au domaine de l'écologie.

Parmi les 35 thèmes, 13 thèmes sont assimilables aux pratiques de modélisation théorique, i.e. ceux qui sont à proximité des ouvrages de référence mathématique et computationnelle (voir Figure 3.2). Dans ce sous-réseau, il y a 12 thèmes plus fortement liés à la modélisation mathématique et un thème, le premier, qui est plus fortement lié à la modélisation computationnelle. Pour cette raison, on met plutôt l'accent sur la modélisation mathématique dans cette section.[4] Les thèmes les plus représentatifs de la modélisation mathématique sont les thèmes 86, 140, 54 et 133 qui portent, respectivement, sur les points d'équilibres, l'espace d'état, la dynamique des populations, et sur les phénomènes d'oscillation et de périodicité dans les systèmes dynamiques. D'autres thèmes, comme le 144, sont notables en ce sens qu'ils sont dominés par un seul terme (p.ex. POPULATION dans le cas du thème 54). On note aussi la présence des thèmes 106 et 144 qui sont sur les taux de croissance et la densité-dépendance. Bien que mathématiques, ces derniers sont davantage méthodologiques, ou fonctionnels, que

---

3. Il est cependant reconnu que le débat autour de la théorie neutraliste de l'évolution est fortement méthodologique. Mais cela s'apparente plus à une discussion en philosophie des sciences que dans la pratique comme telle (Nitecki et Hoffman, 1987)

4. On tient à préciser que plusieurs thèmes sont quand même à proximité de la modélisation computationnelle, notamment les thèmes 133, 140, 116, 48, et 86. Cela dit, dans l'ensemble, ceux-ci restent plus fortement associés à la modélisation mathématique. On revient sur ce problème dans le prochain chapitre, où on soutient que les pratiques de modélisation sont mieux comprises comme des communautés de thèmes.

thématiques.

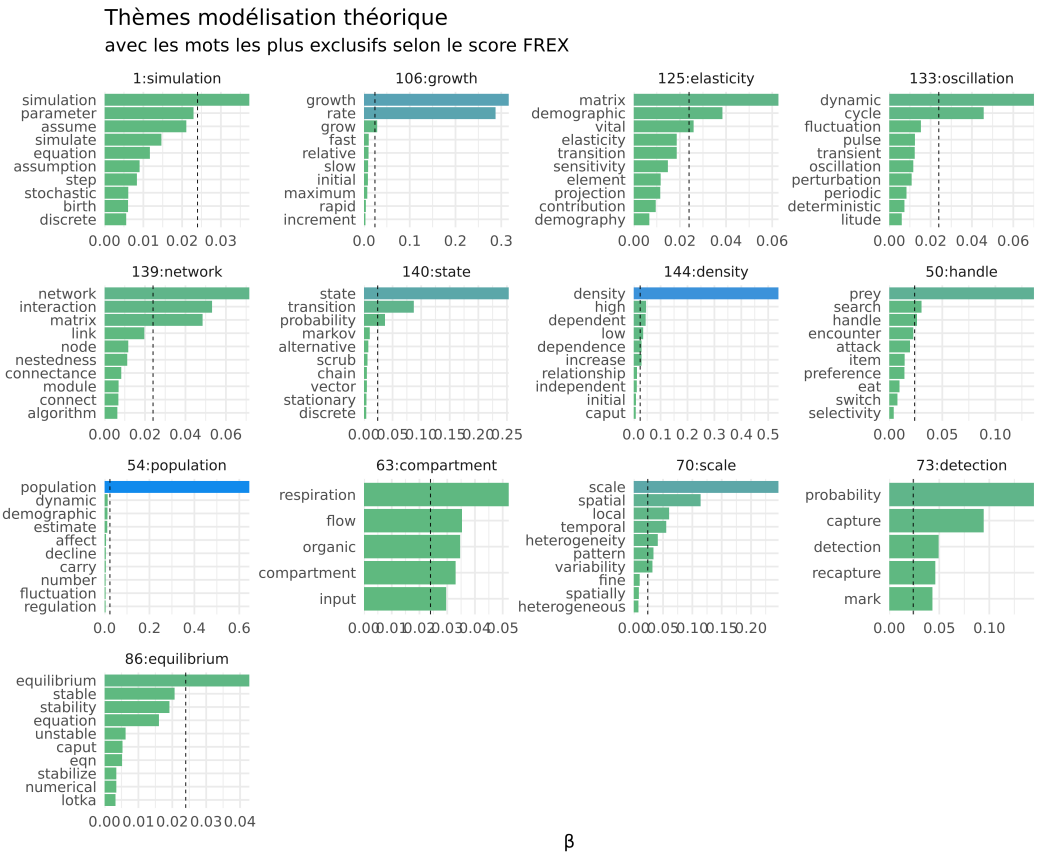


Figure 3.2: Les thèmes identifiés comme en lien avec la modélisation théorique, i.e. mathématique et computationnelle. Ces thèmes ont été identifiés en raison de leur proximité avec les termes représentatifs de l'ouvrage de Otto & Day. Afin de comparer l'importance relative des thèmes, on indique la valeur bêta moyenne du sous-ensemble par la ligne pointillée ($\mu_{\beta\_math} = 0.024$), et on colore les termes en fonction de la force de la valeur bêta.

On note la présence de certains thèmes dont la classification pourrait être remise en cause. En tête de liste, le thème 63 sur les compartiments. Ce thème est particulier en ce qui a trait à la pharmacocinétique. Bien qu'il s'agit d'un domaine d'étude en soi, les modèles de compartiments sont importants en modélisation mathématique parce qu'ils se généralisent à plusieurs systèmes, d'où la décision de garder le thème. Il y a aussi le thème 50 qui à première vue semble davantage en lien avec le domaine de l'écologie que la modélisation mathématique. Cela dit, lorsqu'on examine les articles qui

sont représentatifs de ce thème, on note que ces derniers sont en lien avec les modèles d'optimisation des comportements (en lien avec l'approvisionnement et la prédation), notamment chez les petits invertébrés marins. Après avoir interprété le sens des termes du thème dans leur contexte, et en considérant une similarité significative entre celui-ci et l'ouvrage d'Otto & Day (25%), on décide d'interpréter ce thème comme théorique. Enfin, le thème 125 en lien avec l'élasticité pourrait être interprété comme statistique, mais les termes sont définitivement plus proches sémantiquement des documents mathématiques que statistiques. [5]

---

5. On pourrait s'imaginer que d'avoir choisi d'autres ouvrages de référence comme corpus de validation aurait donné d'autres résultats.

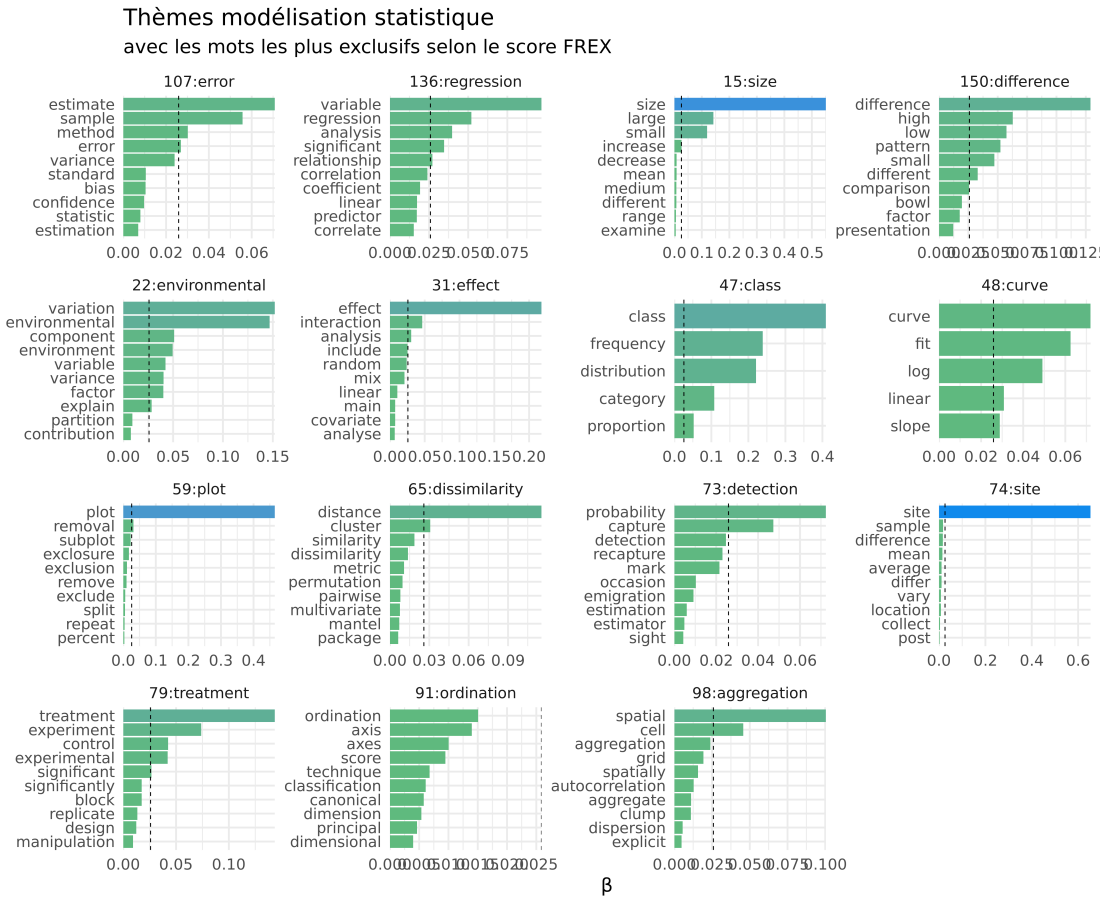


Figure 3.3: Les thèmes identifiés comme en lien avec le domaine empirique, i.e. modélisation statistique, l'estimation et l'expérimental. La figure est identique à celle du sous-ensemble mathématique, à la différence que la moyenne de la valeur bêta est $\mu_{\beta\_stat} = 0.029$

On trouve 15 thèmes qui sont en lien avec la modélisation empirique (43% des 35 thèmes), lesquels semblent se diviser naturellement entre les thèmes portant sur le domaine expérimental et la modélisation statistique (voir Figure 3.3). Les thèmes que l'on interprète comme explicitement en lien avec la modélisation statistique sont les thèmes 136 :REGRESSION, 31 :EFFECT (les modèles mixtes), 48 :CURVE, et le thème 73 :DETECTION (modèles de détection. On classe également les thèmes sur la dissemblance et les distances (65) et l'ordination (91), ainsi que le thème sur l'estimation (107), comme faisant partie des thèmes sur la modélisation statistique. Les thèmes expérimentaux

sont le 79 :TREATMENT, le 74 :SITE, et le 59 :PLOT.[6] Tout comme dans le domaine théorique, on remarque certains thèmes fonctionnels, comme les thèmes 15, 150, et le 22, qui portent, respectivement sur la taille de l'effet, la taille de la différence, et sur la variation environnementale.

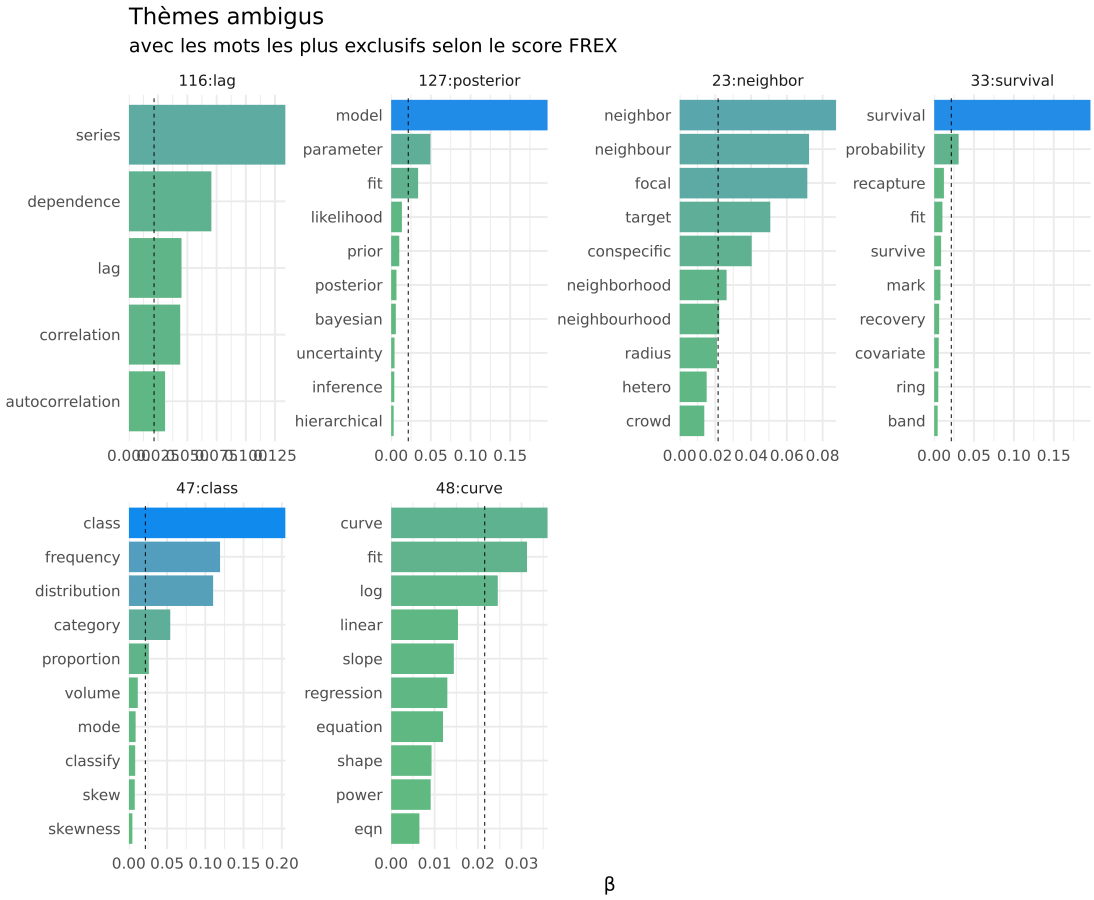


Figure 3.4: Les thèmes qui ont une similarité significative avec les types de modélisation mathématique et statistique. La figure est identique aux précédentes, à la différence que la moyenne de la valeur bêta est $\mu_{\beta\_ambigous} = 0.021$

Certains thèmes sont ambigus, c'est-à-dire que leur similarité sémantique avec les différents types de modélisation est non-négligeable (Figure 3.4). Ces thèmes sont au cœur

6. Celle-ci n'étant pas au centre de notre analyse, nous nous permettons de faire cette distinction de manière informelle.

de notre problème, car ils remettent en question une distinction claire entre le domaine théorique et empirique. Par exemple, le thème 116, qui semble être sur les processus autorégressifs, comprend la présence du modèle de Ricker qui est un modèle bien connu en dynamique des populations nommé après Bill Ricker. Bien que celui soit à proximité de la modélisation empirique, on va voir que celui-ci se trouve à proximité des modèles théoriques lorsqu'on considère les sous-réseaux de thèmes. On attribue la présence des thèmes 47 :CLASS et 23 :neighbor comme ambigus à l'ouvrage de Whitlock & Schluter, qui se concentre définitivement sur les statistiques traditionnelles. On suppose qu'un autre livre contemporain à la jonction de l'apprentissage machine et des statistiques aurait classifié ceux-ci comme empiriques. Compte tenu de ce qui a été dit dans le premier chapitre de la thèse, on considère également que le thème 127 sur les modèles bayésiens est susceptible d'être plus hétérogène qu'il n'y paraît.

On note que la proportion de thèmes liés à la modélisation est assez similaire parmi les revues choisies. Étant donné que le paramètre γ représente la proportion d'un thème par revue,une manière de calculer proportion de thèmes liés à la modélisation de chaque type est de sommer γ de tous les articles pour chaque revue, puis de normaliser. Ce faisant, on obtient les résultats suivants :

Tableau 3.1: Résumé des différentes pratiques de modélisation. La proportion du poids des thèmes en lien avec la modélisation varie de 34% pour *Journal of Animal Ecology* à 30% pour la revue *Journal of Ecology*.

Titre revue	**Tot γ contenu** (%)	**Tot γ maths** (%)	**Tot γ stats** (%)	**Tot γ ambigus** (%)
Eco. Mono	70	10	16	4
Ecology	67	10	19	4
An. Ecology	66	10	18	6
Jn of Eco.	70	7	20	3

Malgré la présence de seulement 34 thèmes en lien avec la modélisation, on note que

le total γ pour les thèmes mathématiques et statistiques gravite plutôt autour de 32% pour chacune des revues. Si l'on considère que les 116 thèmes restants se partagent 70% du poids des thèmes discutés, cela signifie que les thèmes de modélisation sont plus fréquemment discutés que la moyenne (en moyenne 1% du poids des thèmes discutés par article de modélisation contre 0.6%). Or, étant donné qu'on a sélectionné les sections qui traitent des modèles de manière à privilégier le ratio modélisation-contenu en faveur de la modélisation, on attribue cette surreprésentation des thèmes en lien avec la modélisation notamment à notre nettoyage. On remarque aussi une plus faible présence des thèmes mathématiques dans la revue *Journal of Ecology*.

Cela complète notre survol des différents thèmes en lien avec les pratiques de modélisation. Bien que les thèmes soient dans l'ensemble relativement homogènes du point de vue des pratiques de modélisation, on trouve toutefois plusieurs thèmes qui sont ambigus. Cette ambiguïté est l'une des raisons pour lesquelles le niveau des thèmes n'est pas suffisant pour répondre à nos objectifs de recherche. Pour mettre à l'épreuve la vision de Weisberg, il est nécessaire d'esquisser la composition des communautés de thèmes qui composent l'ensemble du territoire.

3.2 Des communautés de thèmes

Une seconde façon de représenter les thèmes du corpus est d'examiner le réseau des plus fortes corrélations entre les thèmes, i.e. quels thèmes tendent à cooccurrer dans les documents. On représente les relations de corrélation des thèmes de trois manières, chacune permettant de mieux représenter l'ensemble du territoire d'intérêt que ce qui a été présenté précédemment. Si tel est le cas qu'il existe une forte séparation entre les pratiques de modélisation théorique (computationnelle, mais surtout mathématique) et empirique (statistique), on s'attend à ce que les thèmes identifiés comme appartenant à l'un et l'autre domaine se retrouvent dans des communautés sémantiques distinctes.

3.2.1 Réseau de corrélation

On examine d'abord les plus fortes relations de corrélation entre les thèmes. Pour avoir une idée de l'émergence des regroupements de thèmes les plus importants, on examine l'évolution de la structure du réseau au fur et à mesure que l'on ajoute plus de relations. Dans la Figure 3.5, on a les 100, 150 et 200 plus fortes corrélations. À noter, on a utilisé les étiquettes données par le score FREX pour mettre l'accent sur les termes typiques propres aux thèmes. [7]

On note tout d'abord que, à 100 relations, le réseau est hautement déconnecté et essentiellement constitué d'îles, interprétées comme des thèmes « inséparables » (Figure 3.5 ; en haut). D'une part, on a des regroupements qui correspondent aux sujets en écologie dont la plausibilité nous donne confiance dans les résultats de la STM. Par exemple, on apprend que le thème 10 :STIGMA, POLLINATION est fréquemment accompagné de celui des 82 :BEES, que le thème de la 46 :PHOTOSYNTHESIS survient toujours avec celui des 135 :LEAVES, et plusieurs autres. De plus, on remarque la présence d'une clique qui est composée d'un sous-réseau de noeuds, à savoir 8 :SPECIES, 94 :RICHNESS, 9 :COMMUNITY, et 104 :ASSEMBLAGE dans laquelle toutes les paires de sommets sont adjacentes. Enfin, il y a des thèmes de nature méthodologique plus proche de nos intérêts.

Toujours à 100 relations, on note que le thème 133 :EQUILIBRIUM est fortement connecté au 86 :OSCILLATION, qui semble représenter des termes en lien à la modélisation mathématique, une triade 31 :EFFECT, MIX, 59 :PLOT et 79 :TREATEMENT, laquelle est en lien avec la modélisation statistique et l'expérimentation (ou un regroupement plus empirique), et finalement une autre triade qui comprend les thèmes 127 :POSTERIOR,BAYESIAN, 107 :ERROR, et 48 :CURVE.

---

7. Rendez vous sur le lien `http://shiny.initiativesnumeriques.org/ecology-corpus-explorer/` et cliquer sur l'onglet EXPLORE TOPIC MODELING / EXPLORE TOPIC NETWORK pour explorer le réseau de chacune des revues et au niveau d'analyse souhaitée

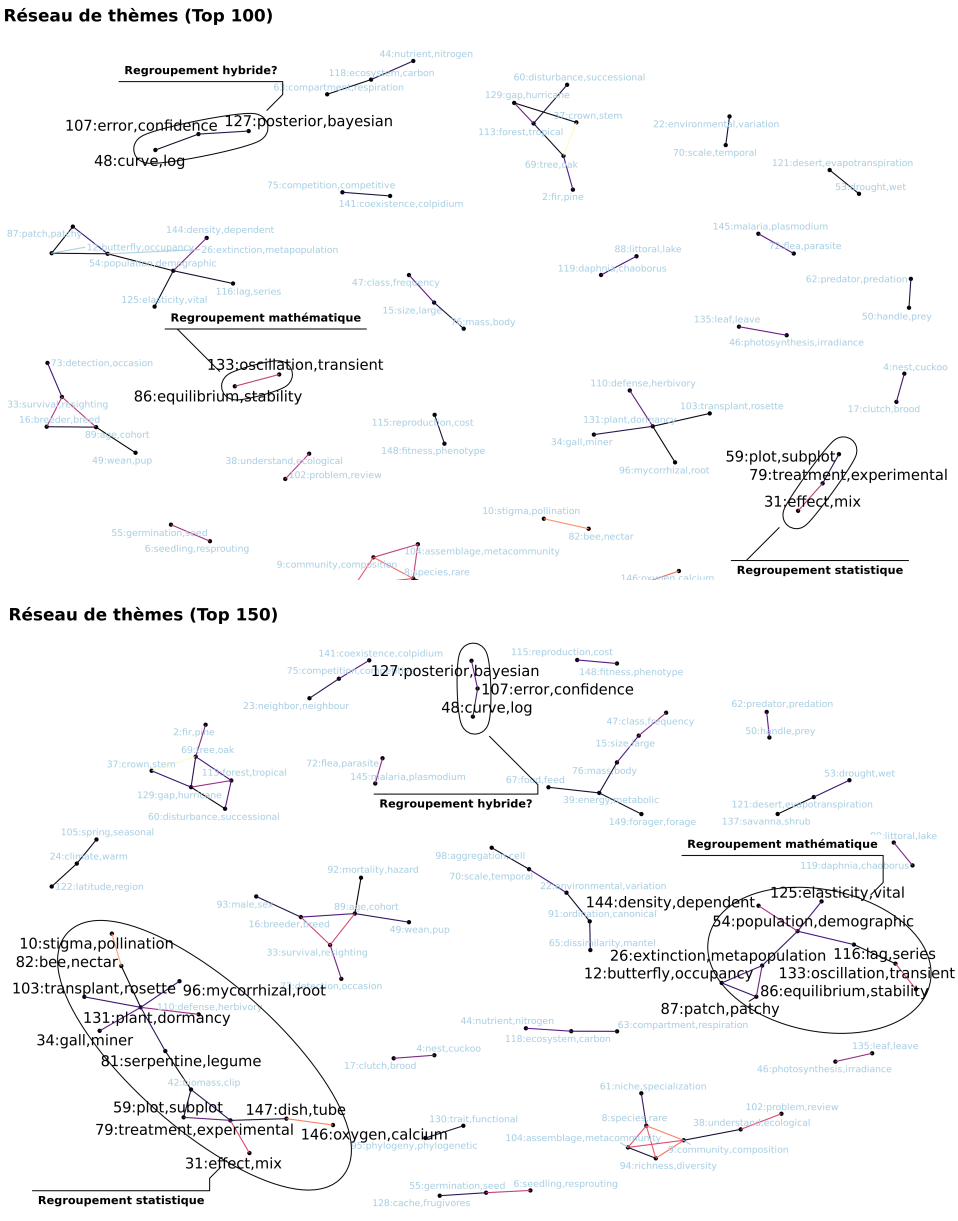


Figure 3.5: Évolution du réseau des plus fortes corrélations. On identifie les regroupements qui sont en lien avec les différents types de modélisation. Lorsque le réseau est composé des 200 relations les plus fortes, on propose que les types de pratiques de modélisation soient mieux compris comme sous-réseaux de thèmes dont l'appartenance au groupe est mixte.

Lorsqu'on considère les 150 relations les plus fortes, on remarque que la dyade de modélisation mathématique a augmenté en nombre (Figure du milieu ; regroupement à droite). En plus de la dyade initiale, on retrouve notamment à proximité le thème 116 :LAG, SERIES, le thème 54 :POPULATION, lequel est accompagné de 114 :DENSITY et de 125 :ELASTICITY. En ce qui concerne le regroupement empirique (regroupement en bas à gauche), celui-ci compte désormais les thèmes 147 :DISH, TUBE et 146 :OXYGEN,CALCIUM, et, d'autre part, le thème 42 :BIOMASS et 81 :SERPENTINE (le mot le plus probant ici est SOIL, et SERPENTINE est un type particulier de SOIL). Ce regroupement est proche du domaine expérimental, et plus particulièrement des expériences dans le domaine de l'écologie végétale.

Finalement, à 200 relations, le regroupement qui contient le thème bayésien se rallie au regroupement identifié comme celui de la modélisation mathématique (Figure 3.5 ; en bas). En plus de la triade de nature bayésienne, on note aussi la présence du thème 1 :SIMULATION,SIMULATE qui vient se rattacher au thème 86 :EQUILIBRIUM. Cela dit, le regroupement mathématique devient lui-même massif, et comprend désormais deux autres grands regroupements de thèmes qui semblent être en lien avec l'étude des écosystèmes (9 :COMMUNITY, 39 :ENERGY, 104 :ASSEMBLAGE, etc) et l'écologie animale (16 :BREEDER, 33 :SURVIVAL, 89 :COHORT, etc). Il n'est pas clair à ce point que le rattachement du thème 127 :POSTERIOR,BAYESIAN au regroupement mathématique est significatif de la pratique hybride.

3.2.2 Clustering hiérarchique

Comme seconde façon de représenter la structure sous-jacente des patrons d'interactions entre les thèmes, on utiliser le clustering hiérarchique agglomératif de Ward. [8] Cette technique regroupe les thèmes qui partagent le même genre de relation de similarité, et

8. Voir l'appendice pour les détails de cette technique de visualisation

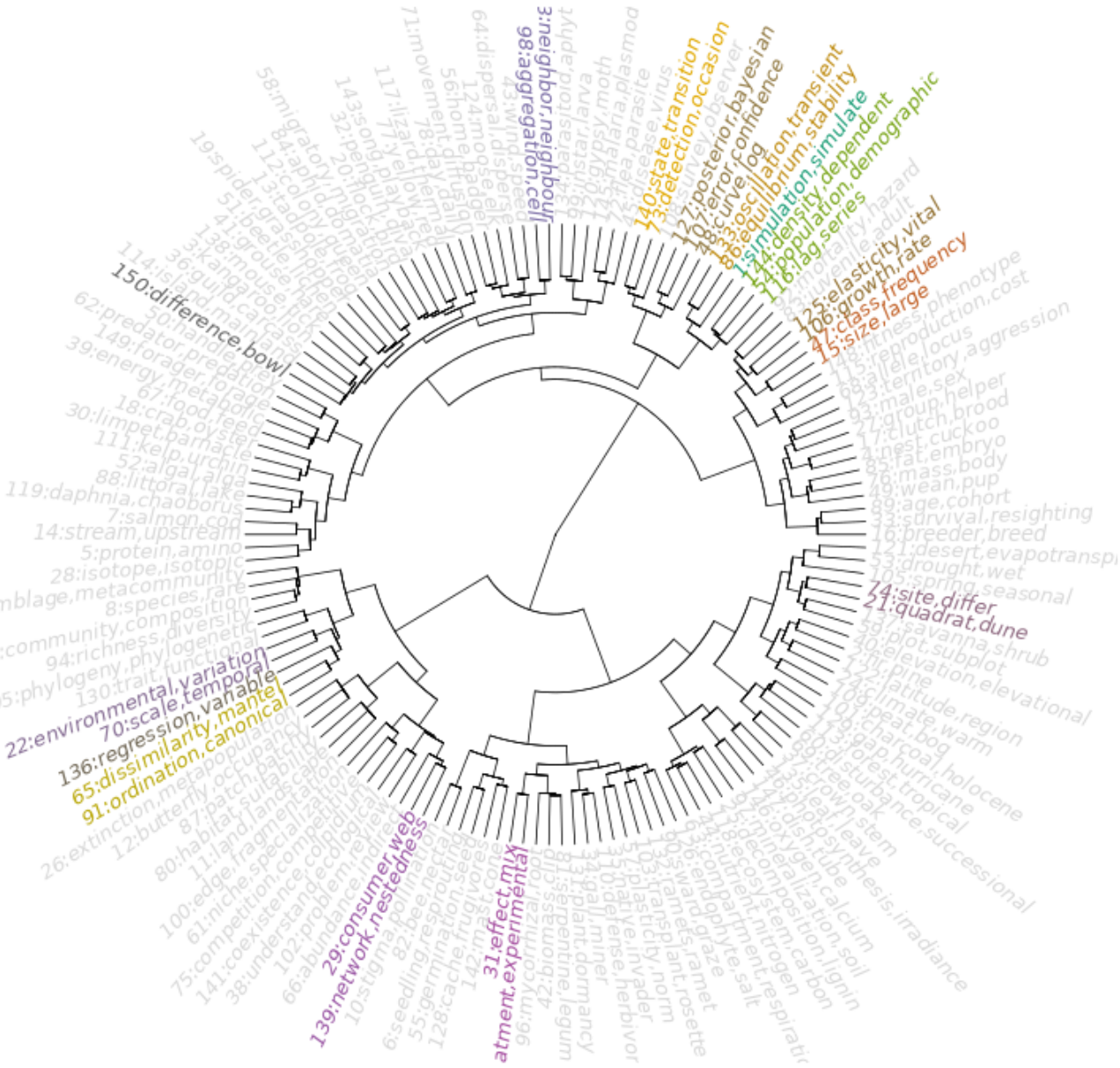


Figure 3.6: Clustering hiérarchique agglomératif avec la méthode de Ward des relations de corrélations entre les thèmes. Les termes liés à la modélisation ont été colorés pour faciliter l'interprétation. On note que l'algorithme regroupe les thèmes associés à la pratique de modélisation mathématique à proximité du sous-réseau hybride, suggérant qu'il existe un nombre important de documents dans le corpus qui mettent en relation ces différents contextes.

En examinant le coin supérieur droit de la Figure 3.6, on constate que l'algorithme regroupe les thèmes liés à la modélisation mathématique (en vert, turquoise, orange brûlé, et brun). Il est intéressant de voir que le thème 127 :POSTERIOR,BAYESIAN, avec le 48 :CURVE,LOG et 107 :ERROR, reste à proximité dans le regroupement adjacent. L'opposition entre les regroupements empiriques et les regroupements théoriques persiste avec les thèmes 31 :EFFECT, 79 :EXPERIMENTAL (en mauve) étant à l'opposé des thèmes que mentionné ci-haut. Au sein du regroupement statistique, on peut distinguer ce qui à trait à l'expérimentation de ce qui est en lien avec la modélisation statistique, à savoir les thèmes 136 :REGRESSION, 91 :ORDINATION, et 22 :ENVIRONMENTAL,VARIATION (en jaune, gris foncé, et violet).

3.2.3 Réduction de dimensionnalité

Enfin, on utilise la projection UMAP, qui a l'avantage d'être hautement interprétable, pour visualiser les relations de corrélation dans un espace à deux dimensions (voir Figure 3.7).[9] Pour faciliter l'interprétation, on a surligné les thèmes identifiés en lien avec les différentes pratiques de modélisation. Les thèmes en dessous des coordonnées (-1,-1) suggèrent, de nouveau, que certains des thèmes plus expérimentaux sont très proches de l'écologie végétale. Par exemple, les thèmes 79 :TREATMENT et 59 :PLOT sont entourés de 135 :LEAF, 96 :MYCORRHIZAL, 40 :ELEVATION. On note que l'espace autour des coordonnées (2,1) est en lien avec le comportement animal (33 :SURVIVAL, 58 :MIGRATORY, 93 :MALE,SEX), alors que l'on retrouve les insectes (72 :FLEA) et d'autres types de relations (10 :STIGMA, POLLINATION, 36 :GUILD, CARCASS) sur le territoire entre les plantes et les animaux (∼0,0). L'algorithme regroupe les thèmes de modélisation théorique dans le haut de la page (1.5, 1.5). Sans surprise, les thèmes portant sur l'oscillation, la simulation, et les POPULATIONS sont quasiment l'un par dessus l'autre, avec ceux de la densité-dépendance, LES POINTS D'ÉQUILIBRES, la croissance à

9. Voir l'appendice pour les détails de cette technique de visualisation.

proximité. Contrairement au clustering hiérarchique, on note que le thème 139 :NETWORK se retrouve à proximité des modèles mathématiques, et à distance égale du thème 127 :POSTERIOR,BAYESIAN.

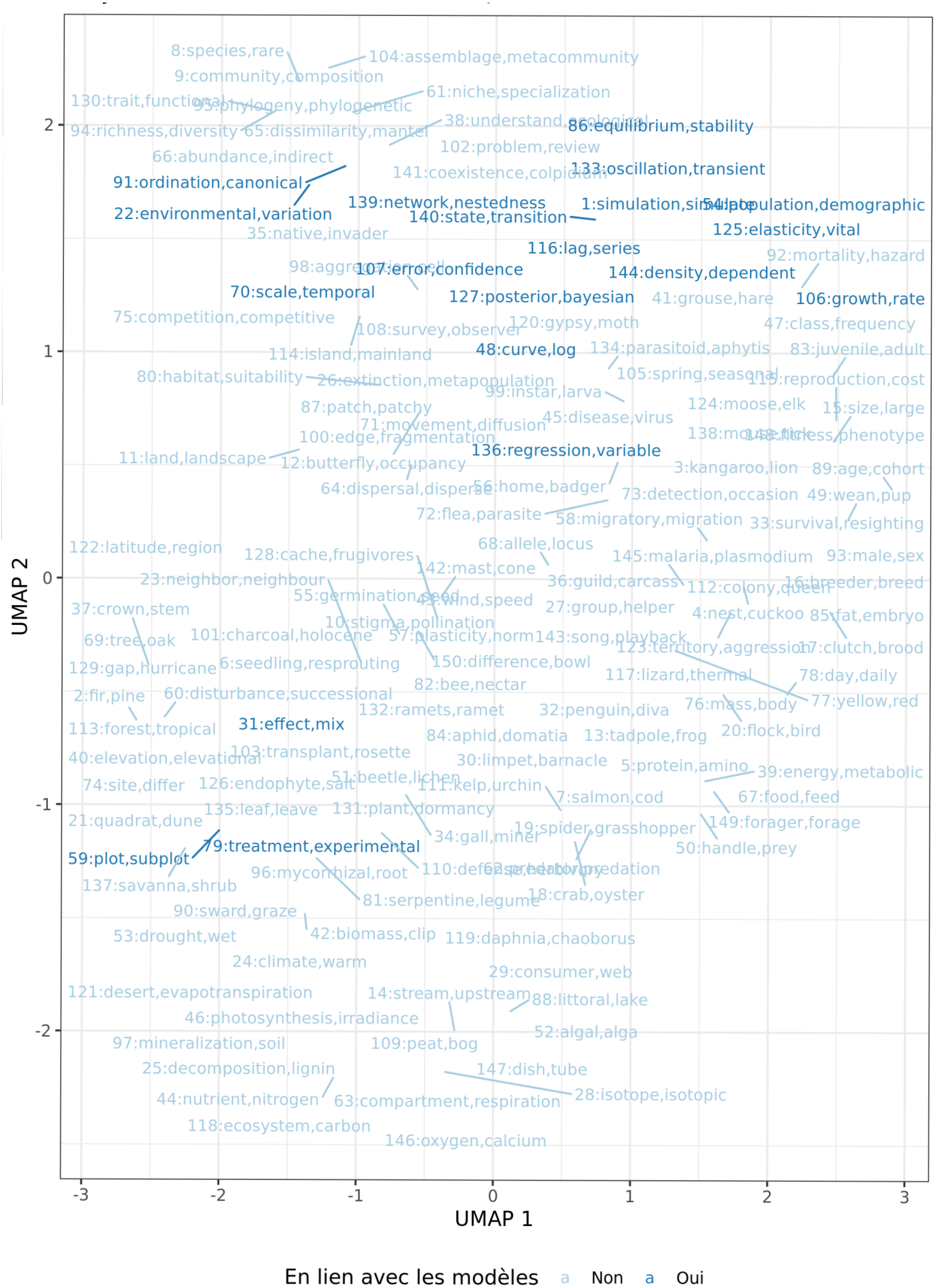
3.3 Comment les thèmes ont-ils évolué au fil du temps ?

Jusqu'à présent, on a représenté les thèmes de manière synchronique. On a vu que, en moyenne, certains thèmes sont plus fréquents que d'autres, que certains thèmes tendent à cooccurrer ensemble, et qu'il est possible de regrouper des thèmes dans de plus grandes communautés qui semblent correspondre aux différentes pratiques de modélisation. Maintenant que l'on a esquissé les contours du territoire, on examine la manière dont les communautés de thèmes ont changé dans le temps. L'évolution des thèmes est importante à considérer parce que le territoire est en flux constant. Il est fort possible que l'hypothèse de Weisberg représente adéquatement les pratiques de modélisation à une époque donnée, mais qu'elle ne garde pas ce statut indéfiniment. En effet, tout comme l'émergence des ordinateurs dans les années 1960 a engendré la pratique de modélisation computationnelle, on s'imagine que la pratique hybride pourrait être le fruit d'une autre révolution.[10] Pour répondre à cette question, on examine la manière dont les thèmes évoluent à l'aide de l'évolution des proportions des thèmes.

La Figure 3.8 (voir ci-dessous) illustre l'évolution des proportions du sous-ensemble de thèmes en lien avec la modélisation par revue. L'axe des ordonnées variant de thème en thème, on utilise la moyenne globale, représentée par la ligne pointillée, pour comparer l'importance relative des thèmes. On voit, par exemple, que le thème 91 :ORDINATION est nettement plus important que le thème 45 parce que la ligne pointillée reste au ras du sol pour le premier et non pour le second. De plus, les thèmes identifiés en lien

10. Il y a des raisons de croire que les progrès des méthodes de Monte-Carlo par chaînes de Markov associées à une plus grande puissance de calcul et une augmentation de la quantité de données pourraient être ce type de révolution attendue(Diaconis, 2008; Kroese *et al.*, 2014).

Figure 3.7: Projection UMAP (avec les étiquettes FREX)



avec les pratiques de modélisation sont pour la plupart au-dessus de la moyenne. En effet, la moyenne globale est de $\gamma_{\text{global}} = 0.66\%$, alors que la moyenne des thèmes de la Figure 3.8 est de $\gamma_{\text{modèle}} = 1.12\%$.

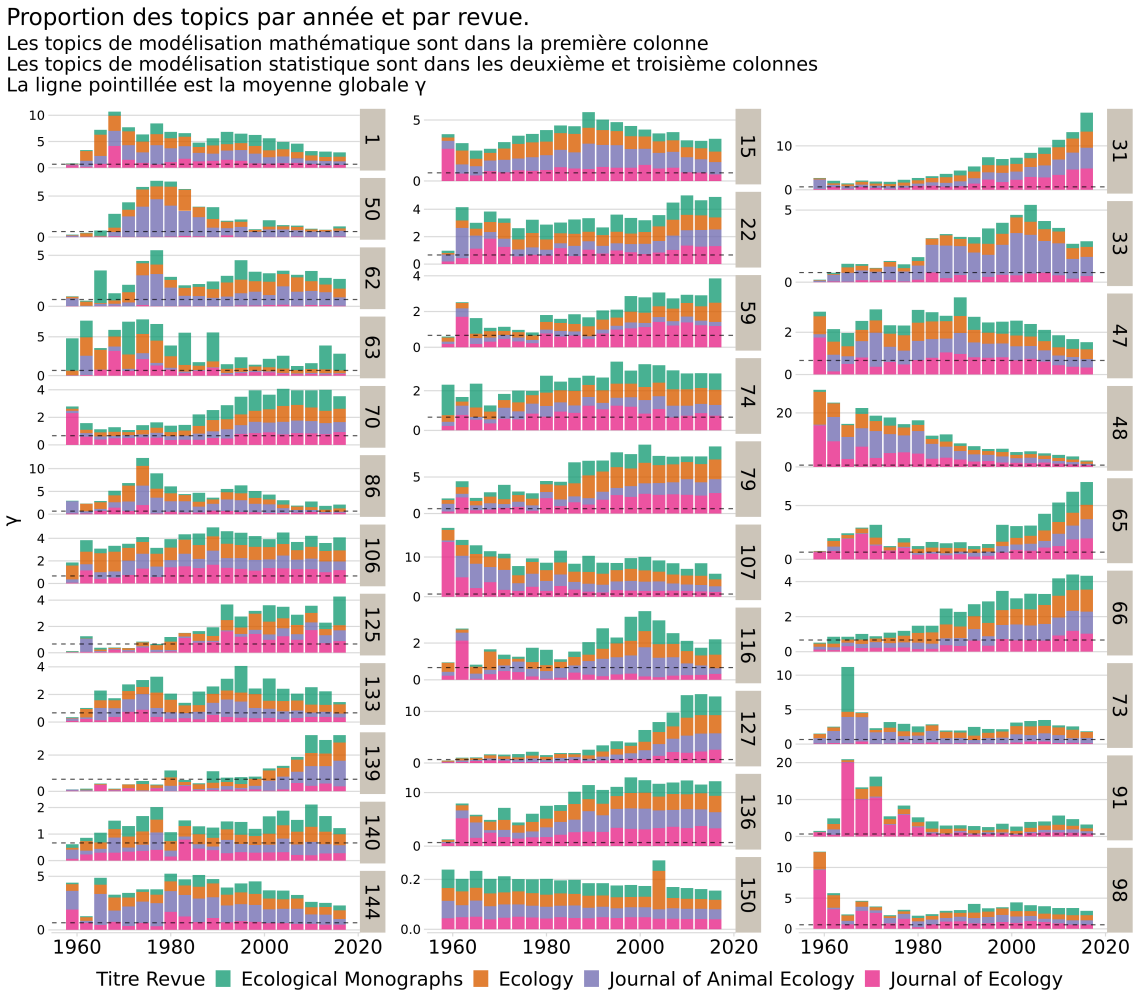


Figure 3.8: Évolution de la proportion des thèmes par années et par revue. Chaque barre représente l'agrégation du paramètre γ sur trois ans. Les barres ne sont pas normalisées pour mettre en évidence la contribution relative de chaque thème, ce qui a pour effet que l'axe vertical varie d'un panneau à l'autre, i.e., certains thèmes sont beaucoup plus fréquents que d'autres. La ligne pointillée, qui est la moyenne globale γ de l'ensemble des thèmes et des revues (indiquées par la couleur des bandes), sert d'ancrage pour s'orienter.

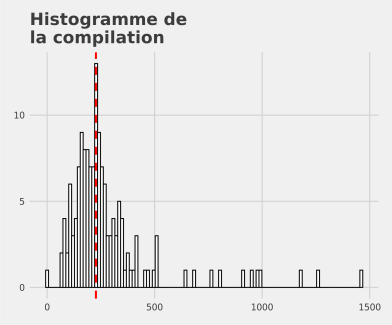
On note quatre grandes tendances dans la Figure 3.8. La popularité d'un thème peut soit augmenter, diminuer, avoir atteint un sommet et ensuite diminuer, ou rester à

peu près la même. Parmi les thèmes qui ont déjà connu leur sommet de popularité, du moins dans le présent corpus, on trouve le thème 91 :ORDINATION, le 86 :EQUILIBRIUM. Les thèmes qui connaissent une chute de popularité sont les thèmes 48 :CURVE, le 107 :ERROR, et le 1 :SIMULATION, alors que les thèmes 139 :NETWORK, 127 :POSTERIOR, 31 :EFFECT,MIX, 136 :REGRESSION connaissent une forte croissance, ainsi que les thèmes 125 :ELASTICITY, 33 :SURVIVAL dans une moindre mesure. Finalement, on a des thèmes plus stables dans le temps comme le 106 :GROWTH, 144 :DENSITY, et le 22 :ENVIRONMENTAL, VARIATION, lesquels sont de nature fonctionnelle.

Aussi, il y a quelques différences notables dans la force des patrons selon les journaux. Par exemple, le thème 59 qui porte sur les mesures contrôlées dans les parcelles est significativement plus élevé dans la revue *Journal of Ecology*, laquelle porte sur l'écologie végétale, que dans le journal *Journal of Animal Ecology*. En contrepartie, le journal *Journal of Ecology* discute moins fréquemment des points d'équilibres (86), des réseaux (139), et, bien évidemment, des modèles de survie et des données de relocalisation (33). Cela dit, bien que les effets varient, on note que la directionnalité des tendances est néanmoins similaire d'un journal à l'autre dans certains cas.

Dans l'ensemble, on note que les thèmes en lien avec les pratiques de modélisation statistiques prennent plus d'importance dans le corpus, au détriment de la modélisation mathématique. Cependant, il faut faire preuve de prudence dans l'interprétation de ce résultat. Un thème qui finit par constituer 10% du corpus, comme le 127 et le 31, est certainement plus hétérogène que l'on voudrait. Par exemple, le thème 127 est difficile à interpréter notamment en raison de la polysémie des termes en lien avec le bayésianisme. Si on compare les termes les plus probables du thème 127 (DATUM, FIT, ESTIMATE) avec les termes les plus exclusifs (LIKELIHOOD, PRIOR, POSTERIOR, HIERARCHICAL), on réalise que la configuration des termes est en lien avec le problème d'estimation. Ce qui n'est pas surprenant étant donné la proximité de l'inférence bayésienne avec la construction de modèles bayésiens.

Pour résoudre cette ambiguïté, il est utile d'examiner les relations entre les thèmes plus attentivement. Pour ce faire, on compile d'abord le nombre de fois où un thème se retrouve parmi les trois thèmes les plus dominants dans un document. Si on examine les 7 thèmes qui reviennent le plus souvent, on obtient :



	N	FREX	PROB
31	914	effect,mix,random	effect,interaction,analysis
38	994	understand,ecological,focus	study,process,ecological
79	1258	treatment,experimental,experiment	treatment,experiment,control
102	1175	problem,review,concept	paper,theory,example
107	973	error,confidence,method	estimate,sample,use
127	958	posterior,bayesian,fit	model,parameter,datum
136	1462	regression,variable,predictor	variable,regression,analysis

Figure 3.9: Résumé des thèmes les plus dominants. À gauche, histogramme du nombre de fois où un thème se retrouve parmi les thèmes les plus dominants. À droite, ce sont les thèmes dont le nombre de d'occurrence est au-delà de $N = 900$

On note que les thèmes identifiés dans le tableau à droite sont nettement en dehors de la normale. On y retrouve notamment le thème 127 :POSTERIOR,BAYESIAN qui se retrouve 958 fois parmi les thèmes les plus fréquents de l'ensemble des documents. Afin de savoir ce qui se cache dans le thème 127 :POSTERIOR,BAYESIAN, on retourne à ses relations de cooccurrence et, dans la prochaine section (Sec. 3.3.1), on fait une incursion dans le texte. Sur les 958 occurrences du thème 127 :POSTERIOR,BAYESIAN, il y en a 722 qui sont partagées avec d'autres thèmes en lien avec les pratiques de modélisation, lesquels sont divisés entre 165 thèmes théoriques (23%) et 557 thèmes empiriques (77%). On décompose à son tour le sous-ensemble de thèmes empiriques entre le thème 107 :ESTIMATION (30% des 557 thèmes empiriques), les thèmes 116, 136, 31, 33, 48, 98 en lien avec les *modèles de régression* (54%), l'analyse multivariée (10%), les thèmes 79, 59 sur le domaine *expérimental* (4%), et le reste qui est de nature fonctionnelle (2%). De même, on divise les relations théoriques entre le thème 1 en lien

avec la *modélisation computationnelle* (35% des 165 thèmes théoriques), et le reste qui est sur la *modélisation mathématique* (65%).

En résumé, le thème 127 :POSTERIOR,BAYESIAN est hétérogène. On peut affirmer que son contexte d'utilisation se divise majoritairement entre les enjeux d'estimation (15%), la modélisation statistique (49%), computationnelle (8%), et mathématique (15%). Pour conclure cette section sur la modélisation thématique, on retourne aux données textuelles pour examiner si les documents qui regroupent les statistiques bayésiennes et la modélisation mathématique correspondent aux documents du corpus de validation.

3.3.1 Une incursion dans le texte

À la lumière des résultats des dernières sections, on soutient que certains documents constitués du thème 127 :POSTERIOR,BAYESIAN sont en fait mieux compris comme une configuration particulière de celui-ci avec des thèmes en lien avec la modélisation théorique. On compare un échantillon textuel de documents identifiés comme ayant une configuration hybride avec les termes prototypiques du corpus de validation.

On trouve un premier article, intitulé *Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts*, qui est un assemblage de thèmes portant sur le bayésien, la simulation, la diffusion et le mouvement (thème 71), l'abondance (thème 66), les données de surveillance (thème 108), et sur les points d'équilibres (thème 86). L'article a été écrit par l'un des auteurs du livre de référence hybride du corpus de validation, Melvin B. Hooten :

> ***Extrait 1 :*** *Ecologists often fit models to survey data to estimate and explain variation in animal abundance. Such models typically require that animal density remains constant across the landscape where sampling is being conducted, a potentially problematic assumption for animals inhabiting dynamic landscapes or otherwise exhibiting considerable spatiotemporal variation in density. We review several concepts from the burgeoning literature on spatiotemporal statistical models, including the nature of the temporal structure (i.e., descriptive or dynamical) and strategies for dimension reduction to promote computational tractability ... Although care must be taken to tailor models to match the study population and survey data available, we argue that hierarchical spatiotemporal statistical models represent a powerful way forward for estimating abundance and explaining variation in the distribution of dynamical populations. (Conn et al, 2015)*

Dans cet article, Hooten discute d'une classe de modèle, les modèles spatio-temporel, qui sont ajustés (*tailor*) au processus sous-jacent de l'abondance des populations. Cet extrait nous donne confiance que la modélisation thématique est en mesure d'identifier la pratique hybride telle que décrite dans le premier chapitre, à savoir comme une configuration particulière de modélisation statistique, mathématique et de connaissances de domaine.

Parmi les auteur.es les plus prolifiques du sous-ensemble d'articles d'intérêts, on a James S Clark qui a écrit 7 articles. Clark est un auteur bien cité pour son livre *Models for Ecological Data* (2007), et auteur de nombreux articles dont *Why environmental scientists are becoming Bayesians*. Dans le présent corpus, il est coautheur sur l'article *Incorporating Multiple Sources of Stochasticity into Dynamic Population Models*, qui se résume ainsi :

> ***Extrait 2 :*** *Alternatively, under the Bayesian paradigm, parameters are viewed as random variables ; the data are used to update one's beliefs about the distribution of the model parameters. The coherence of the Bayesian method provides a straightforward way to account for observation error in addition to process error. The state-space model framework provides a structure for extending time-series models to include both observation and process error. The data are assumed to arise from an unobserved state variable that represents the "true" dynamic process. This underlying variable evolves over time by a process model that explicitly models process error. The model for the relationship between the actual data and the state variable incorporates observation error. Bayesian state-space models with linear structure and normal error distributions allow entirely analytic results we review the relevant expressions. We discuss extensions and alternative approaches for models with nonlinear structure and nonnormal error distributions, including Markov chain Monte Carlo posterior simulation. (Calder et al, 2003)*

On note dans cet extrait l'attention portée à la modélisation de l'erreur, conjointement avec le modèle de processus. Les auteurs y discutent une extension des modèles autorégressifs de dynamiques de population pour inclure l'incertitude causée par l'erreur du processus et par l'erreur d'observation. Cet article regroupe principalement du thème bayésien, avec le thème en lien sur l'espace d'état (140), la simulation (1), l'estimation (107), les modèles autorégressifs (116), les modèles de régression (136), et la dynamique des populations (54). Dans le chapitre sur l'interprétation, on soutient que cet article est un exemple où la taxonomie de Weisberg échoue à représenter cette pratique de modélisation observée en écologie. Au lieu de considérer l'article nécessairement comme théorique ou empirique, on suggère qu'il s'agit d'un de figure où la taxonomie profiterait d'ajouter la pratique hybride.

3.4 Le plongement lexical comme outil d'analyse diachronique complémentaire

Avec la modélisation thématique, on s'intéresse principalement aux thèmes et à leur proportion dans les documents. De ce fait, cette pratique ne représente pas directement les relations sémantiques dans le temps. En délaissant les thèmes au profit du plongement lexical, on peut directement modéliser les relations de proximité sémantique par intervalle de temps discret $t = 1, \ldots, T$. On commence simplement et on examine les mots les plus similaires au terme *model* lors des quatre dernières décennies :

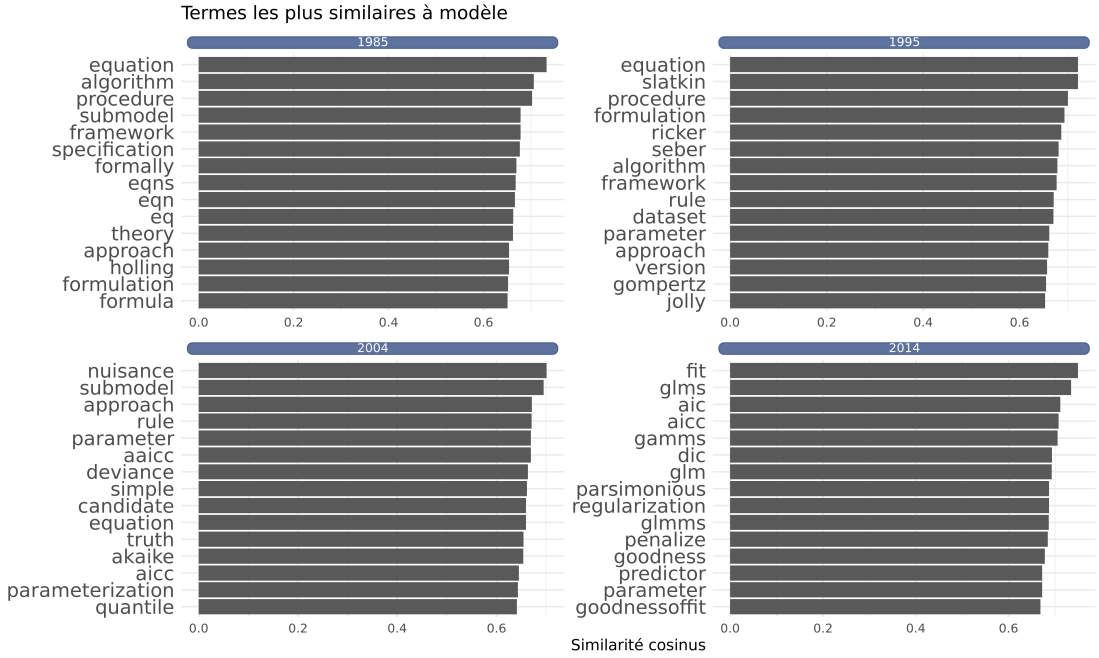


Figure 3.10: Les termes les plus similaires au terme *model* pour quatre années distinctes.

On remarque la manière dont les termes à proximité de *model* ont évolué au cours des dernières décennies, à savoir qu'il y a une tendance des mathématiques vers les statistiques. En effet, en 1985 on trouve que la grande majorité des termes est en lien avec la modélisation mathématique (*equation, holling, theory, formula*), alors qu'en 2014 les termes sont remplacés par des termes en lien avec les statistiques (*fit, glms, gamms*) et la sélection de modèles (*aic, aicc, dic, parsimonious*).

Bien que ces résultats permettent de valider la qualité de l'algorithme à représenter le corpus, ils ne permettent pas de valider si les pratiques hybrides deviennent de plus en plus mathématiques. Pour ce faire, on représente les relations sémantiques sous forme de graphe, où les arêtes sont le lien de similarité, et les noeuds sont les termes d'intérêts. Pour mettre en évidence les différents contextes dans lesquels le terme *model* apparaît, on construit ce réseau à partir des termes les plus similaires à *model*, ainsi qu'à partir des termes les plus similaires à eux.[11] En mettant de côté le terme *model*, que l'on sait déjà connecté à l'ensemble des termes du réseau, on obtient les visualisations de la Figure 3.11. On a choisi de représenter trois années d'intérêts suffisamment éloignés d'un point de vue sémantique pour que les réseaux soient représentatifs de leur époque respective (voir page ci-dessous).

On rappellera que le seuil de similarité à partir duquel les noeuds apparaissent correspond à la médiane de la similarité, et que la taille des étiquettes est proportionnelle à la mesure de la centralité pagerank. On aperçoit dans le graphe de plongements lexicaux des mots appartenant à des regroupements qui sont similaires aux champs sémantiques trouvés par la modélisation thématique. Par exemple, autour des années 1985, le terme *simulation* est relié à l'élasticité, au processus de Monte Carlo, et voisin du regroupement sur les procédures et algorithmes, alors qu'en 2006, le terme a perdu sa proximité avec élasticité, et est désormais relié aux termes *computer*, *baseline*, *iteration*, et *excel*.

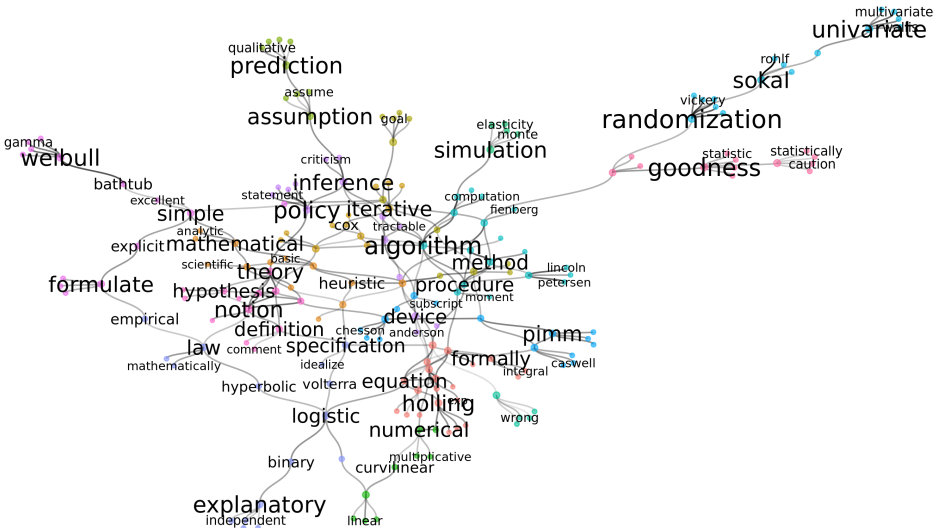
Toujours en 1985, on note une tendance aux pratiques de modélisation théoriques, lesquelles sont distinctes de la modélisation empirique. Plus particulièrement, il y a un regroupement important en lien avec C.S. Holling qui comprend des termes comme *equation*, *logistic*, et *lotka*. Cette visualisation permet aussi de voir la relation intime

---

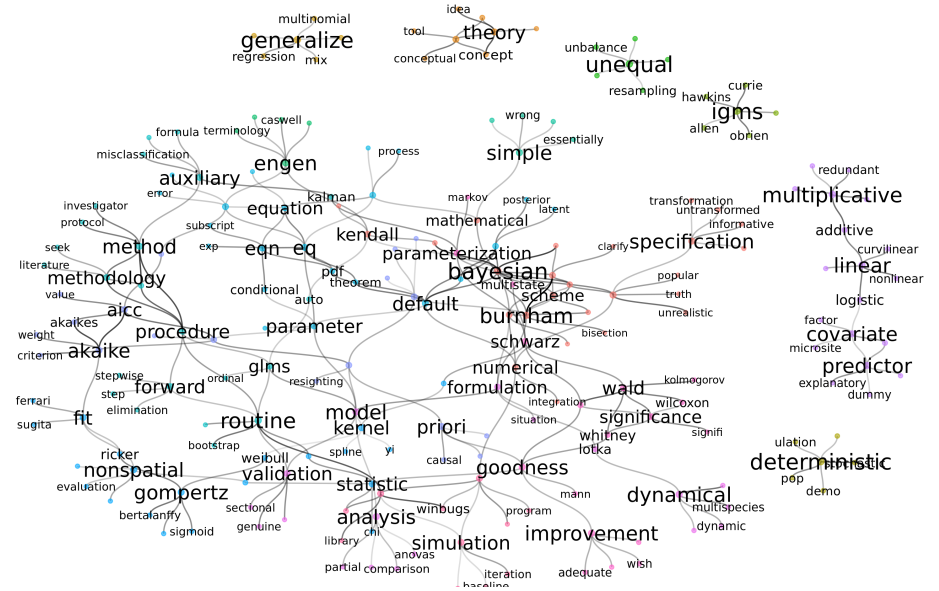
11. Rendez vous sur le lien `http://shiny.initiativesnumeriques.org/ecology-corpus-explorer/` et cliquer sur l'onglet Explore Word embedding / Graph Embedding pour explorer le graphe de plongements lexicaux de chacune des revues et à la fenêtre de temps souhaitée. Dans cette visualisation, on tire avantage de l'interactivité de l'application pour éviter que le réseau soit saturé de texte.

Réseaux des termes les plus similaires au terme "model"

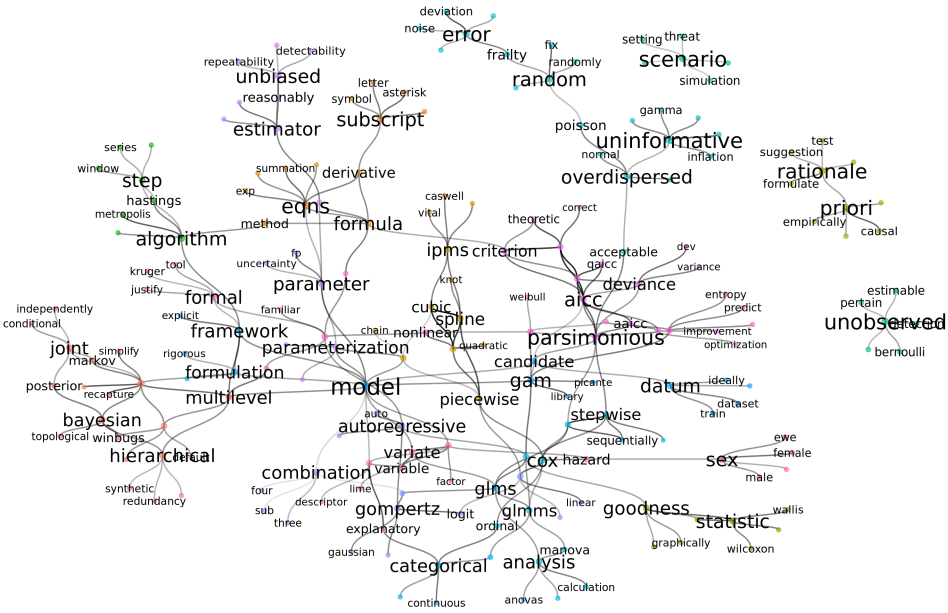
1985 (Taille des noeuds proportionnelle à la centralité pagerank)



2006



2011



entre les notions épistémiques de *concept*, *theory*, et *idea*, et les notions mathématiques en écologie, notamment avec des termes comme *optimality*, *derivation*, et *analytic*. On trouve d'autres regroupements importants portent sur les algorithmes, l'inférence, et la randomisation.

Si on avance de 20 ans dans le futur, en 2006, on note que le réseau est plus mixte. Il y a encore des communautés mathématiques, notamment avec les modèles de Gompertz et Bertalanffy dans le coin inférieur gauche, et le regroupement déconnecté sur la modélisation en dynamique des populations dans le coin droit. Cela dit, on remarque aussi la présence accentuée des enjeux empiriques, tel que défini dans le corpus de validation. On trouve dans le bas du réseau des termes en lien avec les tests statistiques (*analysis*, *anovas*, *goodness*, significance), à gauche la sélection de modèles (*akaike, aicc, criterion*), à droite les modèles de régression (*predictor, linear, additive*). On remarque la présence d'une communauté sémantique bayésienne, mais celle-ci est liée au cadre d'analyse de Kenneth P. Burnham. Burnham a écrit avec Anderon un article très cité en 1998 (plus de 50,000 fois) sur l'utilisation des mesures d'information-théoriques pour la sélection de modèles.

Une façon complémentaire de vérifier la présence de la tendance vers une hybridation est d'examiner dans quelle mesure le graphe de plongement lexicaux du terme bayésien devient lui même de plus en plus mathématique avec le temps (Figure 3.12). Si les termes en lien avec la modélisation empirique diminuent au profit du théorique, cela est une preuve qu'il y a suffisamment d'occurrences de contexte hybrides pour influencer le graphe de plongement lexicaux. On illustre ce réseau deux moments clés, à savoir lors de l'émergence des pratiques bayésiennes dans les années 1990, et en 2012 :

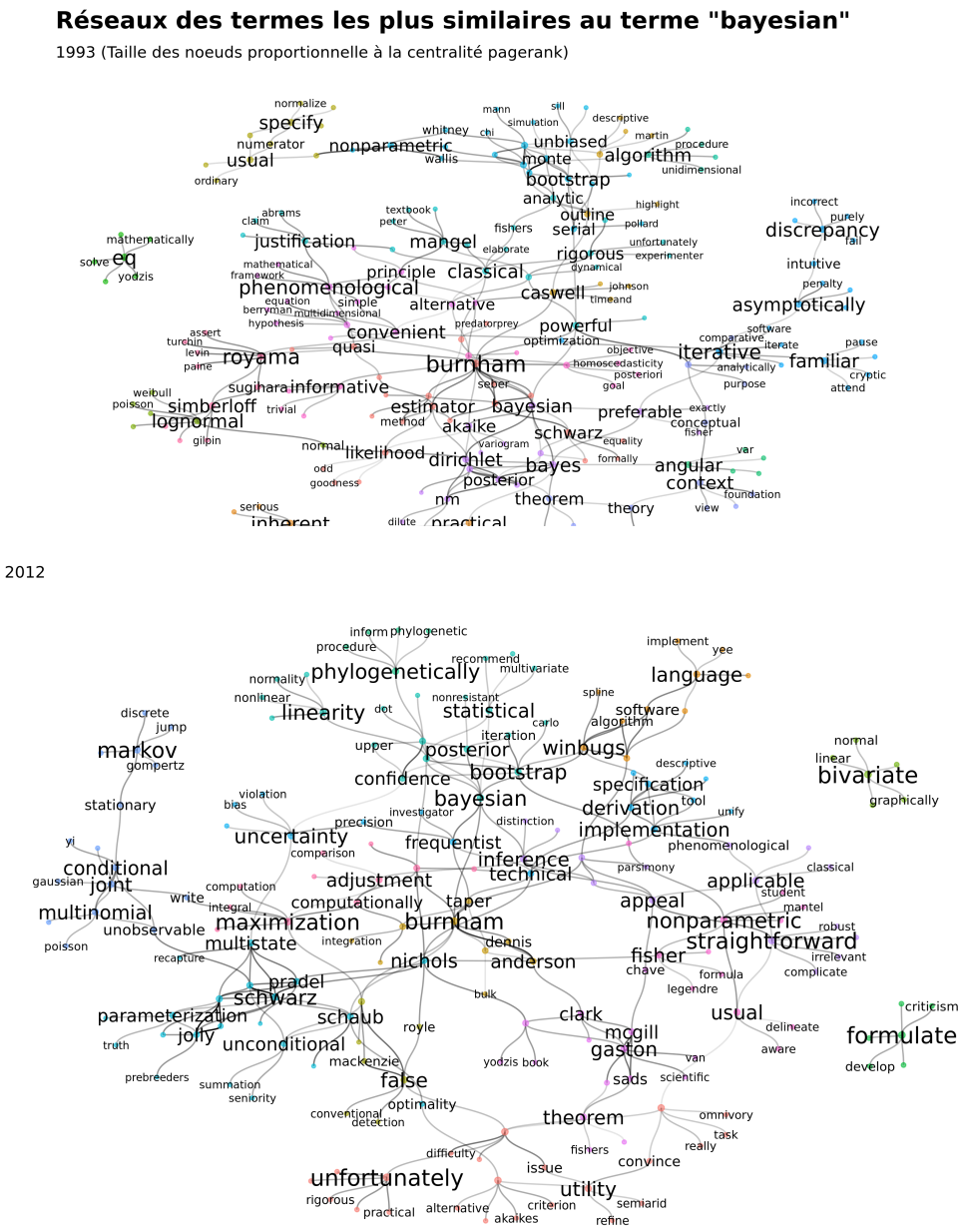


Figure 3.12: Les graphes de plongements lexicaux du terme *bayésien* en 1993 et en 2012.

Dans la figure précédente, le graphe des années 1990 est fortement en lien avec les statistiques. On y retrouve des regroupements sur l'*asymptotic*, le *bootstrap*, plusieurs

communautés en lien avec les distributions statistiques (*lognormal, dirichlet, weibull, normal*). On remarque aussi des termes en lien avec la controverse entourant l'utilisation des modèles bayésiens, i.e., *viewpoint, preferable, context*. En comparaison, le réseau de 2012 est beaucoup plus diversifié. On note des regroupements en lien avec l'analyse phylogénétique (dans le haut du réseau), l'analyse *capture–recapture* avec le modèle de Jolly-Seber (à gauche), ou encore l'utilisation d'un processus de Markov dit de saut discret» (*Jump process*), lequel est utilisé au sein du modèle mathématique de Gompertz.

C'est ce qui complète le résumé de nos principaux résultats obtenus par l'analyse thématique et le plongement lexical. On a entrepris la présentation des résultats en examinant la structure et l'évolution des communautés pour ensuite aller dans les détails jusqu'à retourner aux textes bruts. Dans le prochain chapitre, on fait le chemin inverse. On soutient que la taxonomie de Weisberg ne parvient pas à représenter adéquatement le type de pratique de modélisation à l'oeuvre dans l'article de Calder *et al*, avant de retourner au niveau des structures communautés, et de contextualiser ce type d'articles dans la cartographie.

CHAPITRE IV

INTERPRÉTATION DES RÉSULTATS ET DISCUSSION

Dans l'introduction, on a proposé que la contribution des humanités numériques à la philosophie des sciences s'apparente à la validation des modèles en apprentissage automatique. Si c'est le cas, nos résultats devraient nous permettre de dire si la taxonomie de Weisberg se généralise à l'ensemble des données de notre corpus. Qu'entend-on par généraliser ? Weisberg affirme que les types de modèles se distinguent par leur structure et leur interprétation. Par exemple, les modèles computationnels sont distincts des modèles mathématiques parce que les règles de transition, ou les algorithmes, sont ce qui explique le phénomène d'intérêt (Weisberg, 2013, p.32). On propose que le modèle de Weisberg ne se généralise pas si les catégories proposées par son modèle ne sont pas représentatives, ou déforment, certaines pratiques observées dans le corpus.

Pour commencer, on revient sur l'article de Calder *et al* (2003), reconnu comme hybride par nos méthodes, pour expliquer en quoi celui-ci serait difficile à catégoriser sous la taxonomie de Weisber (Sec. 4.1). On soutient que le modèle de Weisberg n'est pas en mesure de représenter adéquatement ce type d'articles, car leur cœur explicatif correspond à la pratique hybride décrite dans la deuxième partie du premier chapitre. Une fois cela fait, on remet les articles hybrides en perspective dans le contexte plus large de notre cartographie ; on discute de la place relative des thèmes de ces articles au sein des sous-réseaux d'interactions thématiques (point de vue structurel des communautés de thèmes ; Sec. 4.2) et des tendances évolutionnaires des sous-réseaux d'intérêts (Sec. 4.3).

En conclusion, on note les limites de notre approche et on discute la nécessité d'inclure la pratique hybride dans la taxonomie de Weisberg (Sec. 4.4).

4.1 Lecture rapprochée d'un article hybride

Parmi les articles identifiés comme ayant un champ sémantique hybride, l'article de Calder *et al.* (2003) est un exemple où le modèle de Weisberg ne parvient pas à se généraliser. Dans l'extrait proposé à la Section 3.3.1, on remarque que le cadre bayésien proposé dans l'article ne se limite pas à une description des patrons de données. Les auteurs introduisent un modèle qui contient le processus de génération de données dans son ensemble. En d'autres termes, bien qu'ils supposent la présence d'un processus dynamique latent, la contribution de leur modèle consiste à relier le phénomène observé (le processus observationnel) au système latent (le modèle de processus). De manière analogue à la distinction de Weisberg entre la modélisation computationnelle et mathématique, le coeur explicatif de cet article n'est pas la modélisation de la transition du système de processus selon une équation différentielle, mais plutôt la modélisation du processus dynamique en prenant en compte les différentes sources d'incertitudes du système d'intérêt. De même, cet extrait ne se réduit pas non plus à de la modélisation statistique, telle que définie précédemment (Sec. 1.2.1). Il y a des suppositions précises et informées par la théorie dans le modèle, notamment sous forme de variables d'états.

De plus, l'article de Calder *et al.* est un cas de figure intéressant qui nous permet de faire le pont entre la lecture rapprochée du texte et la structure latente découverte par l'analyse thématique. Dans le présent cas, l'article est largement dominé par le thème 127 :POSTERIOR,BAYESIAN (51%), suivi des thèmes 107 :ERROR (17%), 140 :STATE,TRANSITION (8%), 1 :SIMULATION (6%), et 116 :LAG (5%). Tel qu'observé avec le réseau des corrélations les plus fortes (Sec.3.2.1), ce type de configuration de thèmes ressemble plus à la modélisation dite théorique qu'à la modélisation empirique. On réitère le fait que cette configuration de thème est surprenante du point de vue de

la taxonomie de Weisberg. En effet, Weisberg suppose que la modélisation statistique n'a pas sa place pour comprendre les pratiques de modélisation théorique, ce qui a pour effet de nous forcer à classifier cet article soit comme empirique ou soit comme théorique. Étant donné que l'extrait ne correspond pas à la modélisation mathématique, ou computationnelle, la prédiction de son modèle serait que la pratique de modélisation sous-jacente à cet extrait doit être empirique. L'alternative est de tout simplement reconnaître la place de cette pratique pour ce qu'elle est, à savoir une pratique hybride. Cela dit, étant donné la proximité de l'assemblage ci-dessus avec le domaine théorique, on pense malgré tout que la pratique hybride mérite d'être considérée comme théorique quand vient le temps de comparer les pratiques de modélisation théorique.

On trouve 135 articles qui sont similaires à la configuration de thèmes retrouvée dans l'article de Calder *et al*, à savoir les thèmes 127 et 116 sont au moins à plus de 5% et des thèmes mathématiques au moins à 2%. Ces articles représentent une présence non négligeable de la pratique hybride dans notre corpus. On remet en perspective ces articles dans la cartographie en examinant nos résultats en lien avec la structure des communautés thématiques.

4.2 La structure des interactions thématiques

Dans la Section 3.2.1, on a vu que le réseau de corrélation met en évidence un regroupement de thèmes théoriques (mathématiques et computationnels) et empiriques, i.e. les thèmes en lien avec les statistiques expérimentales et les modèles de régression traditionnels. Ainsi, comme le suppose Weisberg, à un certain niveau d'agrégation on distingue aisément les pratiques théoriques de celles qui sont empiriques. Si on se fie au corpus de validation, le regroupement empirique est composé de thèmes méthodologiques tels que le 79 :TREATMENT, 31 :EFFECT, MIX, 136 :REGRESSION, 59 :PLOT, SUBPLOT, lesquels sont étroitement liés à des thèmes en écologie végétale (p.ex. les thèmes 147 :DISH, TUBE, 81 :SERPENTINE, 6 :SEEDLING RESPROUTING).

De plus, on constate que cette distinction entre le domaine empirique et mathématique semble robuste, car elle se retrouve peu importe la technique utilisée. En effet, ces deux grandes communautés sont sous-jacentes à l'embranchement le plus général dans le clustering hiérarchique, et elles semblent prédictives de l'emplacement des thèmes dans la projection UMAP.

Or, si on examine attentivement les relations à proximité du regroupement mathématique, la distinction théorique-empirique semble insatisfaisante. Le regroupement mathématique contient suffisamment de thèmes ambigus sémantiquement proches de la modélisation empirique pour émettre l'hypothèse que la distinction entre les deux semble s'éroder. On propose que nos résultats supportent l'existence d'une pratique hybride au même titre qu'ils supportent l'existence de la modélisation computationnelle. Dans les deux cas, la modélisation hybride et computationnelle est un sous-réseau de la modélisation mathématique dont la configuration rend le type de pratique distincte. On suggère que la présence du thème bayésien (127), accompagné d'autres thèmes discutés dans l'article de Calder, peut être interprétée comme signalant l'émergence de la pratique hybride dont le signal est encore faible, mais présent.

4.2.1 Zoom in sur l'approche hybride

Si on retourne aux relations de corrélation (voir Figure 3.5), on note que le thème 127 :POSTERIOR,BAYESIAN est corrélé positivement avec des thèmes de modélisation statistique (107 :ERROR, 91 :ORDINATION, 31 :EFFECT, 136 :RÉGRESSION), théorique (140 :STATE, 1 :SIMULATION), et d'autres termes ambigus (116 :LAG, 73 :DETECTION). Un examen plus approfondi des relations de corrélation révèle également que le thème 127 : posterior, bayesian est corrélé négativement avec les thèmes "expérimentaux" du regroupement empirique 147 : plat, 60 : perturbation, 132 : ramets. On note que les thèmes qui tendent le plus souvent à cooccurrer avec le thème 127 sont les thèmes en lien avec l'estimation (107) et LAG (116), lesquels, si on se réfère à la Figure 3.3, portent

sur les séries temporelles, l'autocorrélation, et les processus stochastiques.

Le sous-réseau composé de thèmes ambigus et mathématique est mieux compris sous la perspective d'une pratique hybride, dans laquelle les auteurs modélisent le processus biologique et le processus d'observation et font des suppositions informées par la théorie à propos des deux composantes. Par exemple, on a déjà mentionné l'ambiguïté du thème 116 qui partage des liens à la fois avec les thèmes mathématiques et statistiques. En effet, si on observe les thèmes avec lesquels le thème 116 cooccurre le plus souvent parmi le top 3 des thèmes de tous les articles, on trouve qu'il apparaît 40% des fois avec des thèmes mathématiques contre 60% avec des thèmes statistiques. On note que parmi les thèmes statistiques, presque la moitié (44%) sont les 107 :ERROR et 127 :POSTERIOR,BAYESIAN. Parmi les articles qui sont composés des thèmes 116 :LAG, 127 :POSTERIOR,BAYESIAN, et des thèmes mathématiques, on trouve des articles typiques d'une pratique hybride, notamment celui de Calder et al. discuté précédemment.

De plus, nos résultats suggèrent aussi que les articles identifiés représentent une présence non négligeable de la pratique hybride. On interprète les résultats du clustering hiérarchique et de la projection UMAP comme un indice que le signal de cette configuration hybride est suffisamment fort pour que les deux méthodes puissent le capter. Il est important de se rappeler que les deux méthodes partagent l'objectif de regrouper des thèmes qui sont similaires dans leurs relations de corrélation, mais les deux méthodes remplissent cet objectif de manière totalement différente. Ainsi, la proximité entre le thème 127 et la pratique de la modélisation théorique suggère que ce signal est relativement robuste.

En somme, les relations de corrélation permettent d'identifier une structure, ou une configuration de thèmes, qui semble correspondre à la structure des modèles hybrides proposés dans le premier chapitre. Les résultats permettent d'affirmer que cette configuration dite hybride ne fait pas partie du regroupement que l'on a associé au domaine

empirique. Cela dit, il faut se rappeler que le thème 107 sur l'estimation n'est jamais bien loin du thème 116 et 127. Considérant cela, on peut se demander s'il est possible que l'assemblage observé ne soit pas expliqué par l'émergence de la pratique hybride, mais plutôt par la pratique de confronter les modèles théoriques aux données. Plus généralement, est-ce que la structure identifiée est le résultat d'une discussion sur les modèles théoriques et sur ses implications empiriques, ou vice versa, sans que celle-ci soit une véritable configuration hybride ?

4.2.2 L'estimation n'est pas la pratique hybride

L'un des grands avantages de la cartographie à grande échelle est que celle-ci permet de d'abord considérer un grand corpus à l'aide d'algorithmes, puis de retourner à une lecture rapprochée sur un sous-ensemble d'intérêt. Ainsi, la cartographie nous a permis de trouver réduire le nombre d'articles liés à la pratique hybride de 14,000 à seulement 135 articles. Cette réduction importante du nombre d'articles provient de la manière que l'on a construit le sous-ensemble d'intérêt, à savoir en ne gardant que les articles dont les 3 thèmes les plus importants sont identifiés comme faisant partie des thèmes portant sur la modélisation (donc le document doit être déjà dominé par les méthodes), puis que le sous-réseau des thèmes de ces articles respecte la configuration de la pratique hybride. Bien que cette procédure soit restrictive, elle a l'avantage de produire un sous-ensemble d'articles qui peuvent par la suite facilement être examinés pour vérifier si leur composition s'apparente à un assemblage hybride, ou si celle-ci reflète des discussions indépendantes sur le théorique et l'empirique.

Parmi les 135 articles, on trouve que l'article intitulé *Statistical Methods to Correct for Observation Error in a Density-Independent Population Model* ne correspond pas à la pratique hybride. En effet, il s'agit d'un article qui porte sur le problème d'estimation de l'erreur d'observation, et non pas sur la pratique de construction d'un modèle hybride. On note toutefois que dans cet article, le thème 107 :ERROR est très dominant à 65% de

la proportion totale de l'article. Si on contrôle pour la présence forte du thème 107 pour le sous-réseau d'intérêt, on estime que le reste des articles est similaire à nos attentes vis-à-vis la pratique hybride, telles que présentées précédemment. Parmi les articles examinés, on retrouve notamment les documents par Hooten et Clark discutés dans la Section 3.3.1. Bien que ce résultat soit a posteriori, on rappelle le cadre exploratoire de la présente cartographie qui est d'identifier si la piste des modèles hybrides mérite d'être poursuivie éventuellement.

4.3 Dynamique des communautés thématiques

La temporalité des thèmes est une autre manière de distinguer la pratique hybride des autres types de modélisation. On se doute que la pratique hybride est beaucoup plus récente que la celle de la confrontation les modèles théoriques aux données. Déjà dans les années soixante, on trouve des articles comme ceux de Holling (1966) où les modèles mathématiques sont mis à l'épreuve par leur capacité à s'ajuster aux données. Si on retourne à la Figure 3.8 sur l'évolution de la proportion des thèmes, on observe que le thème 107 :ERROR sur l'estimation est fortement présent dès le début et tend à diminuer dans le corpus au fil du temps. Lorsqu'on observe les dates de publication du sous-ensemble identifié, au contraire, celles-ci correspondent à la fenêtre temporelle de l'émergence de la pratique hybride, i.e. du début des années 1990 jusqu'à aujourd'hui.

Pour renforcer le résultat précédent, on note que les résultats des graphes de plongements lexicaux du terme *model* racontent une histoire similaire. En examinant la Figure 3.11, on s'aperçoit que le contexte bayésien est absent des années 1980. Lorsque celui-ci apparaît autour de 2006, il est intimement lié à la sélection de modèles de K.P. Burnham (Anderson *et al.*, 1998). À ce stade, il n'est pas évident que le contexte bayésien se détache du sens statistique au profit des regroupements de mots en lien avec la pratique de modélisation mathématique. Autour des années 2011, on voit que le terme *bayesian* se distingue du contexte de la sélection de modèles, et ne se trouve pas particulièrement

à proximité des regroupements en lien avec les statistiques. Bien qu'il s'agisse seulement d'une preuve indirecte, ce résultat suggère que le contexte bayésien, autour de 2011, est davantage en lien avec la construction de modèles bayésiens. Cette tendance à la construction de modèles, associée entre autres à l'apparition de modèles probabilistes basés sur des principes premiers, est également confirmée par l'examen de l'évolution du graphe de plongements lexicaux du terme *bayesian*.

Dans la Figure 3.12, on note qu'il y a plusieurs regroupements similaires au terme *bayesian* qui sont en lien avec l'estimation. Plus particulièrement, on trouve *estimator* et *asymptotically*, lesquels sont absents 20 ans plus tard. À la lumière de ce qui a été vu avec la modélisation thématique, on interprète de nouveau ce résultat comme une réduction des contextes bayésiens exclusivement liés à l'inférence bayésienne, en faveur des problématiques liées aux modèles bayésiens hiérarchiques, parmi lesquels on retrouve le contexte hybride. Ainsi, pris conjointement avec le résultat de la modélisation thématique, on obtient que le contexte bayésien se rapproche de plus en plus d'un contexte théorique, et qu'il ne soit pas particulièrement proche du domaine empirique/expérimental, tel que défini par Weisberg. Ce résultat fait écho à la proposition de Michael Betancourt (2019) selon laquelle l'utilisation du Théorème de Bayes ne nous rend pas bayésiens aujourd'hui : ce qui nous rend bayésiens, c'est la quantification de l'incertitude avec le langage des probabilités, à laquelle on ajoute la caractérisation du processus latent par une composante déterministe pour obtenir une pratique hybride.

4.4 Les limites

Est-ce que les résultats de la cartographie sont suffisants pour privilégier l'hypothèse d'une pratique hybride de modélisation au détriment de la thèse de Weisberg à l'effet qu'il n'existerait que trois types de modélisation (n'incluant pas la pratique hybride) ? Avant de répondre à cette question, on précise les limitations du mémoire. On revient aussi sur la question de la contribution des humanités numériques aux enjeux tradition-

nellement philosophiques.

Une première « limite » de l'analyse s'apparente à ce que décrit l'écrivain Borges (1941) dans sa nouvelle « Le Jardin aux sentiers qui bifurquent ». Étant donné le nombre important de choix qui ont été faits pour analyser le corpus, lesquels se trouvent à différentes étapes du processus—sélection des revues, nettoyage, prétraitement, granularité de l'analyse, méthodes utilisées—on reste prudent sur les conclusions du présent travail. Bien qu'on ait justifié chacun des choix pris aux différents embranchements, le phénomène des sentiers qui bifurquent stipule qu'un trop grand degré de liberté dans l'analyse mène à des problèmes de comparaison multiples, où l'on compare différents résultats qui dépendent des différents découpages des données. C'est pourquoi on met ici « limite » entre guillemets, car il n'est pas évident que cette situation constitue une limite en soi. Dans l'ensemble, on considère que l'approche utilisée pour cartographier le territoire de la modélisation en écologie est sensible au phénomène de bifurcation, mais que nous avons fait de notre mieux pour faire preuve de transparence et justifier chaque branche. Une piste de recherche intéressante serait de formaliser ce problème et d'utiliser une fonction objective pour quantifier l'effet de nos différents choix.

Un enjeu étroitement lié à cette première limite est le défi d'analyser rigoureusement nos données qui sont massives [1] et hétérogènes. Ainsi, on pense que nos résultats sont valables pour le niveau d'analyse choisi et pour les quatre revues d'intérêts. Dans l'ensemble, il y a une configuration de communautés de thèmes qui est compatible avec l'émergence d'une pratique hybride. Cela dit, nos résultats varient lorsque l'on considère les revues indépendamment les unes des autres. Par exemple, bien qu'observée globalement, la configuration hybride est moins importante si l'on considère seulement la revue *Journal of Animal Ecology*. Malheureusement, une analyse rigoureuse de cette variation entre le niveau global et local est hors la portée du présent mémoire.

---

1. Du moins par rapport à ce qui est fait habituellement en philosophie

Une seconde limite réside dans l'évaluation de la modélisation thématique et du plongement lexical, qui sont toutes deux des méthodes non supervisées. Contrairement aux méthodes supervisées, les méthodes non supervisées requièrent un ensemble de pratiques de validation pour s'assurer que les résultats obtenus sont utiles. Grimmer et Stewart écrivent, « *To validate the output of an unsupervised method, scholars must combine experimental, substantive, and statistical evidence to demonstrate that the measures are as conceptually valid as measures from an equivalent supervised model. Similar to unsupervised methods, validating ideological scaling requires numerous and substance-based evaluation* » (Grimmer et Stewart, 2013). En d'autres mots, la validité des résultats des méthodes non supervisées est sensible aux connaissances du domaine des examinateurs. De ce fait, on considère que les résultats du mémoire bénéficieraient d'une évaluation itérative de la part d'écologues et d'experts sur la modélisation. Tout comme on cherche à éliminer les biais des modèles en utilisant plusieurs modèles indépendants, il serait nécessaire d'inclure dans notre enquête empirique plusieurs examinateurs indépendants qui, idéalement, arriveraient aux mêmes résultats.

Si l'on met de côté les limites plus méthodologiques, on note que la polysémie des termes reste un enjeu important pour établir la validité d'une enquête basée sur la modélisation thématique. On a vu que certains thèmes d'intérêt ont fortement gagné en popularité dans les dernières années, par exemple le thème 127 :POSTERIOR, BAYESIAN ou le thème 139 :NETWORK qui portent respectivement sur les statistiques bayésiennes et sur les réseaux. Cela dit, on a également vu que ces thèmes sont notoires en ce qu'ils ont des termes très polysémiques, ce qui rend difficile de savoir quelle signification des termes est responsable de la croissance du thème. Par exemple, on peut se demander, d'une part, dans quelle mesure les termes bayésiens sont responsables de la montée en popularité du thème 127, et, d'autre part quel sens des termes *bayesian*, *posterior*, *prior* contribue le plus à la montée en popularité du thème 127 ? Bien qu'on a partiellement abordé cette question avec le plongement lexical, en modélisant directement les différents contextes

du terme bayésien, il n'en reste pas moins que la polysémie des termes constitue un obstacle pour cartographier les différentes pratiques de modélisation. [2]

Si on revient à la définition des pratiques de modélisation présentée dans le premier chapitre, on se rappellera que les pratiques de modélisation correspondent à une combinaison de structure de modèle et d'interprétation. Il n'est pas évident que dans la présente étude les méthodes utilisées permettent de dégager l'interprétation des scientifiques, à savoir la description, la fonction d'interprétation, la portée attendue du modèle, et les critères de fidélités. Par exemple, dans l'extrait de Hooten (Sec. 3.3.1), on reconnaît que la ligne est mince entre la modélisation statistique qui s'intéresse à la dynamique des populations (i.e. les processus autorégressifs) et la modélisation hybride. En effet, on pense qu'il s'agit d'un exemple où la structure du modèle ne suffit pas à elle seule à distinguer complètement les deux types de pratiques et qu'il soit nécessaire de faire ressortir l'interprétation des scientifiques. On propose qu'une manière d'aller plus loin dans cette direction serait d'inclure des métadonnées supplémentaires dans le modèle de la STM. En effet, l'interprétation des modèles étant hautement culturelle, une piste serait de modéliser la relation entre les liens d'affiliations et les configurations que l'on suppose correspondre à la pratique hybride.

En plus des limitations au niveau des méthodes, il est possible que l'interprétation des modèles soit tout simplement absente des textes des revues sélectionnées. En effet, on ne croit pas que les revues choisies encouragent les scientifiques à expliciter la manière dont ils interprètent leurs modèles. Dans ce cas, c'est un exemple intéressant où la présence de plus de données du même type ne permet pas de résoudre le problème. Sinon, il est aussi possible que l'analyse du langage naturel ne soit pas un outil adéquat pour représenter cet aspect de la modélisation, qui, par définition, est implicite chez les scientifiques. En résumé, on reconnaît la présence de limites importantes dans ce que

---

2. Par ailleurs, on note que la modélisation de la polysémie des mots en général constitue l'une des frontières actuelles de l'analyse du langage naturel.

l'on peut affirmer sur la présence d'une pratique hybride, qui se distingue notamment par son interprétation, dans le corpus en écologie.

4.5 Doit-on ajouter la pratique hybride au modèle de Weisberg ?

À la lumière de nos résultats, on croit utile de reconnaître une place à la pratique hybride comme une pratique distincte à l'intersection de la modélisation théorique et empirique. En esquissant les contours des différentes pratiques de modélisation en écologie, on a vu que les communautés de thèmes permettent de représenter les types de modèles retrouvés en écologie, i.e., les modèles théoriques (mathématiques et computationnels) et empiriques (statistiques). Si on comprend que les pratiques de modélisation donnent lieu à des configurations particulières d'un ensemble de thèmes lorsque ces pratiques sont exposées dans les articles scientifiques, on a des raisons de croire que la pratique hybride est tout autant un sous-réseau mathématique que la pratique computationnelle. On a vu que le clustering hiérarchique et la projection UMAP ont tous deux mis la modélisation bayésienne à proximité des thèmes mathématiques tels que les points d'équilibres, l'oscillation, et les espaces d'états. Une lecture rapprochée a également permis de constater que les techniques computationnelles utilisées ont permis de détecter avec succès des articles correspondant à la description de la pratique hybride articulée dans le premier chapitre. Enfin, on a pris soin de lever l'ambiguïté du thème 127 :POSTERIOR, BAYESIAN sur le contexte bayésien avec la technique du plongement lexical pour s'assurer que les sens des différents termes d'intérêt correspondent aux sens supposés.

Or, compte tenu des limites énoncées dans la section précédente, on ne pense pas que les résultats sont suffisants pour prouver hors de tout doute l'existence de la pratique hybride. En d'autres mots, bien que l'on croit que nos résultats démontrent l'utilité de supposer la présence d'une pratique hybride pour expliquer les données observées, on ne va pas jusqu'à dire que les résultats permettent de rejeter l'idée selon laquelle trois types de modèles sont suffisants pour capturer l'ensemble des pratiques de modélisation en

sciences (écologie) aujourd'hui. Dans l'ensemble, il est possible d'expliquer les résultats en ne recourant qu'à la pratique de la modélisation mathématique et computationnelle, puis d'interpréter le sous-ensemble des thèmes 127 :POSTERIOR, BAYESIAN, 107 :ESTIMATION et 116 : AUTOREGRESSION comme trop éloignés pour être considérés comme faisant partie de l'espace théorique. Par exemple, on pourrait s'imaginer dans le résultat du clustering hiérarchique qu'au lieu de considérer le thème 127 :POSTERIOR, BAYESIAN comme faisant partie du regroupement voisin de celui des thèmes mathématiques, on insiste plutôt sur le fait qu'il ne fait pas partie dudit regroupement mathématique. Cette interprétation fait écho aux limites de l'utilisation de techniques non supervisées et de l'absence de critères rigoureux et prédictifs pour déterminer quels regroupements représentent le mieux le phénomène d'intérêt.

L'une des contributions du mémoire est de fournir des évidences que les types de modèles ne sont pas isomorphes aux thèmes, i.e. un type de modèle n'est pas représenté par un thème correspondant. Par exemple, bien qu'il n'y ait qu'un thème exclusivement sur la simulation (thème 1) on ne croit pas que la modélisation computationnelle se réduise à celui-ci. Au lieu, on soutient que les pratiques de modélisation correspondent des sous-réseaux de thèmes, comme présenté dans les résultats sur les relations de corrélation. Plus précisément, la modélisation mathématique passe par la configuration du réseau de thèmes. Dès lors, montrer l'existence d'une pratique de modélisation n'est pas sans rappeler des questions en écologie des communautés où l'on cherche à quantifier la composition d'une communauté. Dans notre cas, on cherche à savoir si la composition observée est suffisamment distincte pour privilégier le modèle d'une pratique hybride.

4.6 Retour sur la contribution des humanités numériques à la philosophie

Les outils des humanités numériques nous ont permis d'appréhender de nouvelles façons de poser et de répondre à des questions philosophiques. Considérant que la philosophie des sciences contemporaine s'intéresse à synthétiser les différentes pratiques scientifiques

à un niveau épistémologique, les humanités numériques ouvrent la porte à des analyses à grande échelle. On maintient que la combinaison d'une lecture rapprochée et d'une lecture distante est un atout lorsqu'il vient le temps d'opérationnaliser de valider les propositions synthétiques de la philosophie. En effet, le cadre conceptuel de Weisberg a permis de faire sens d'un grand corpus en écologie.

On réitère toutefois que l'absence de représentation de l'interprétation des scientifiques de leurs modèles correspond à une limite de l'analyse textuelle des revues étudiées. À nos yeux, il s'agit d'une raison évidente pour laquelle une combinaison avec l'analyse qualitative des données, qui n'est pas toujours idéale pour répondre à certaines questions plus précises, reste nécessaire. Rappelons que la pratique hybride, caractérisée par la structure du modèle, laquelle comprend le système latent et le processus d'observation, et son interprétation, vise à modéliser le processus de génération de données. Néanmoins, on n'interprète pas cette absence d'interprétation comme une preuve que l'interprétation n'est pas à l'œuvre dans la pratique hybride, mais plutôt que les revues choisies ne permettent tout simplement pas de représenter cet aspect de la modélisation. Dans ce cas, cette idée reste dépendante d'une lecture rapprochée de textes et de la capacité à démontrer ce point avec les outils de la logique.

CONCLUSION

Afin de mettre à l'épreuve le modèle de Weisberg, une esquisse des pratiques de modélisation en écologie a été réalisée. Cette esquisse offre un aperçu de la production scientifique en lien avec la modélisation de quatre revues en écologie sur soixante ans, à savoir de Levins, pionnier de la modélisation en écologie, à aujourd'hui. L'objectif était, d'une part, de savoir si le modèle de Weisberg, qui postule trois grands types de pratiques de modélisation, se généralise à un ensemble de données non-observées, et, d'autre part, de mieux comprendre le rôle des humanités numériques pour des questionnements philosophiques.

Un premier apport, de nature méthodologique, est d'avoir d'avoir étudié la manière de faire le pont entre la philosophie des sciences et les humanités numériques, à savoir que nous avons exploré un grand corpus à l'aide de différentes méthodes computationnelles dans le but de répondre à des questions d'intérêt pour la philosophie de la modélisation. La supposition clé est que les contextes dans lesquels les scientifiques discutent de leurs modèles sont révélateurs de la manière dont ils ou elles pratiquent la modélisation. En examinant ces différents contextes d'utilisation des modèles à l'aide de la modélisation thématique et du plongement lexical, il a été possible de valider, au moins en partie, les propositions au niveau épistémique de Weisberg. Plus particulièrement, on a montré que la configuration des structures des communautés de thèmes en combinaison avec le plongement lexical à l'échelle des mots est utile pour approximer la structure des modèles utilisés en écologie.

Ainsi, l'une des principales contributions du mémoire est d'avoir quantifié l'évolution de l'utilisation des modèles mathématiques, computationnels, et empiriques dans un

domaine précis des sciences. On a notamment quantifié la proportion des thèmes qui sont en lien avec l'écologie et ceux qui sont méthodologiques, parmi lesquels on retrouve de grandes communautés qui correspondent au domaine empirique et théorique de la modélisation. On a argumenté que des configurations de thèmes permettent de représenter les différents modèles discutés par Weisberg. En visualisant l'ensemble des relations de corrélation à l'aide d'un réseau des plus fortes corrélations (à différents niveaux de granularité), d'un clustering hiérarchique, et la projection UMAP, on a identifié une possible configuration hybride qui relie le thème bayésien à des thèmes théoriques. Le graph embedding a permis de s'assurer que l'évolution des contextes de certains termes prend bien le sens suggéré par l'analyse de la modélisation thématique.

On arrive toutefois à des résultats mitigés concernant la possibilité de privilégier l'hypothèse d'une *pratique* hybride au détriment des pratiques suggérées par le modèle de Weisberg. On a vu dans le premier chapitre du mémoire que la modélisation ne se réduit pas aux modèles utilisés. Avec Weisberg, on accepte que les pratiques de modélisation soient une combinaison des différentes structures des modèles et de l'interprétation de ceux-ci par les scientifiques. On a aussi vu dans le premier chapitre que l'existence d'une pratiques de modélisation hybride dépend de l'utilisation des modèles hybrides, qui contiennent à la fois une composante déterministe et probabiliste, et qui consistent en une interprétation distincte de la cible de modélisation. Plus particulièrement, la pratiques de modélisation hybride suppose que la cible est une combinaison du processus latent et du processus observationnel. Or, il semble que les résultats ne permettent pas de représenter explicitement l'interprétation des scientifiques des modèles.

Concrètement, une dernière contribution de ce mémoire est le corpus lui-même. Bien qu'essentiellement invisible, on souligne que le nettoyage d'un grand corpus d'intérêt pour la philosophie des sciences est en soi une contribution. Bien qu'on ait posé certaines questions précises au sujet de la modélisation, le corpus reste disponible pour d'autres types d'enquêtes. Par exemple, au lieu de se concentrer sur les catégories des

pratiques de modélisation, une avenue intéressante aurait été de répertorier et faire la généalogie de la grande diversité de modèles retrouvés en écologie. Cette étude serait sans aucun doute complémentaire à produire une cartographie à différentes échelles du territoire. Dans tous les cas, les résultats présentés restent une petite fraction de toutes les analyses possibles, lesquelles sont rendues en partie disponibles dans l'application interactive mise en ligne. On espère que cette contribution illustre la manière dont les humanités numériques pourront dans le futur stimuler de nouvelles questions de recherche, notamment en philosophie des sciences.

Pour conclure, malgré des résultats mitigés pour la quantification de l'approche hybride dans son ensemble, on croit tout de même avoir montré, d'une part, que les propositions épistémiques de la philosophie des sciences gagnent à être étudiées, et peut-être validées, par les méthodes d'humanités numériques, et d'autre part que le cadre théorique de la philosophie des sciences permet de mieux explorer et de donner un sens à un grand volume d'articles scientifiques. Plus précisément, la philosophie des sciences façonne des catégories que l'on peut par la suite utiliser pour interpréter les résultats d'algorithmes non supervisés. On se satisfait d'avoir su quantifier une structure possible des modèles hybrides, et on laisse à plus tard la quantification de la pratiques de modélisation hybride dans son ensemble.

APPENDICE A: NOTES SUR LE NETTOYAGE DES DONNÉES

Qu'est-ce qu'on entend par des fichiers texte bruités ?

Aujourd'hui encore, la technologie OCR redonne des fichiers texte très bruités. Et plus nous remontons dans le temps, plus le bruit est exacerbé en raison de la moindre qualité des PDFs numérisés. Qu'est-ce que nous entendons par bruit exactement ? Voici un extrait typique :

> *Rams with horns describing at least four- fifths of a curl ("legal" males) were hunted from late August through October. In 1972-1981, removals of December 1999 2541</page><page sequence="4">ANNE LOISON ET AL. adult females kept the June population at 95-110 sheep (Jorgenson et al. 1993b). There was an average of 175 sheep in 1982-1997. Sheep River (50 ? N, 114 ? W, el- evation 1420 to 2550 m) included the seasonal ranges of a migratory bighorn population.*

Dans l'extrait, on remarque les éléments du PDF qui ont été convertis avec le reste du texte, comme le nombre de pages et le nom de l'autrice. Il y a des caractères spéciaux qui ont été mal compris par l'OCR. C'est le cas pour "°" qui est devenu " ?". Un dernier problème apparent dans cet extrait est la manière dont les mots sont séparés en fin de ligne.

Une fois nettoyé, l'extrait ci-dessus prendra la forme suivante :

> *"ram" "horn" "describing" "fourfifth" "curl" "legal" "male" "hunted" "late" "august" "october" "removals" "december" "adult" "female" "june" "population" "sheep" "jorgenson" "average" "sheep" "sheep" "river" "elevation" "included" "seasonal" "range" "migratory" "bighorn" "population"*

Bien que cette représentation semble moins informative que la précédente, il faut se rappeler que la représentation en sac de mots (*bags of words*), qui est utilisée pour notre analyse de analyse thématique, fait fi de l'ordre des mots. L'expérience sur le terrain révèle que les tokens sont plus que souvent suffisants pour faire ressortir la structure latente dans le texte, étant donné les suppositions présentes dans l'analyse thématique.

Analyse exploratoire du corpus d'écologie

Afin de se faire une idée sur les particularités de notre volumineux corpus en écologie et de la qualité du nettoyage, on propose faire une brève analyse exploratoire des données. On utilise une variante du tf-idf, à savoir *le ratio log-odds pondéré*. Similaire au tf-idf, le ratio log-odds calcul le ratio du nombre d'occurrences d'un mot dans deux documents A et B, divisé par le nombre total de mots des documents respectifs, puis compare les deux documents en prenant le logarithme du ratio des valeurs calculées. Dans l'implémentation utilisée (Silge 2020), cette valeur est ensuite pondérée avec un prior « non-informatif » provenant de la Loi de Dirichlet pour prendre en compte la variabilité d'échantillonnage (voir *appendice B* pour les détails techniques). L'un des avantages du ratio log-odds pondéré sur le tf-idf est que si tous les auteurs utilisent les mêmes termes, ceux-ci n'obtiennent pas un score nul, comme c'est le cas avec tf-idf. Au lieu de cela, les termes sont pondérés par leur utilisation. On examine d'abord le *ratio log-odds pondéré* par décennie pour chaque revue (voir Figure 4.2).

Malgré notre nettoyage qui favorise le contenu sémantique en lien avec la modélisation, on remarque que les termes les plus représentatifs des revues sont pour la plupart des termes en lien avec l'écologie. Il est bon de savoir que la revue *Journal of Ecology* s'intéresse surtout à l'écologie des plantes, alors que la thématique de *Journal of Animal Ecology* est le comportement animal (les deux revues étant publiées par la *British Ecological Society*). On remarque également dans la revue *Ecological Monographs* la présence de « macArthur » et « equations » dans les années soixante.

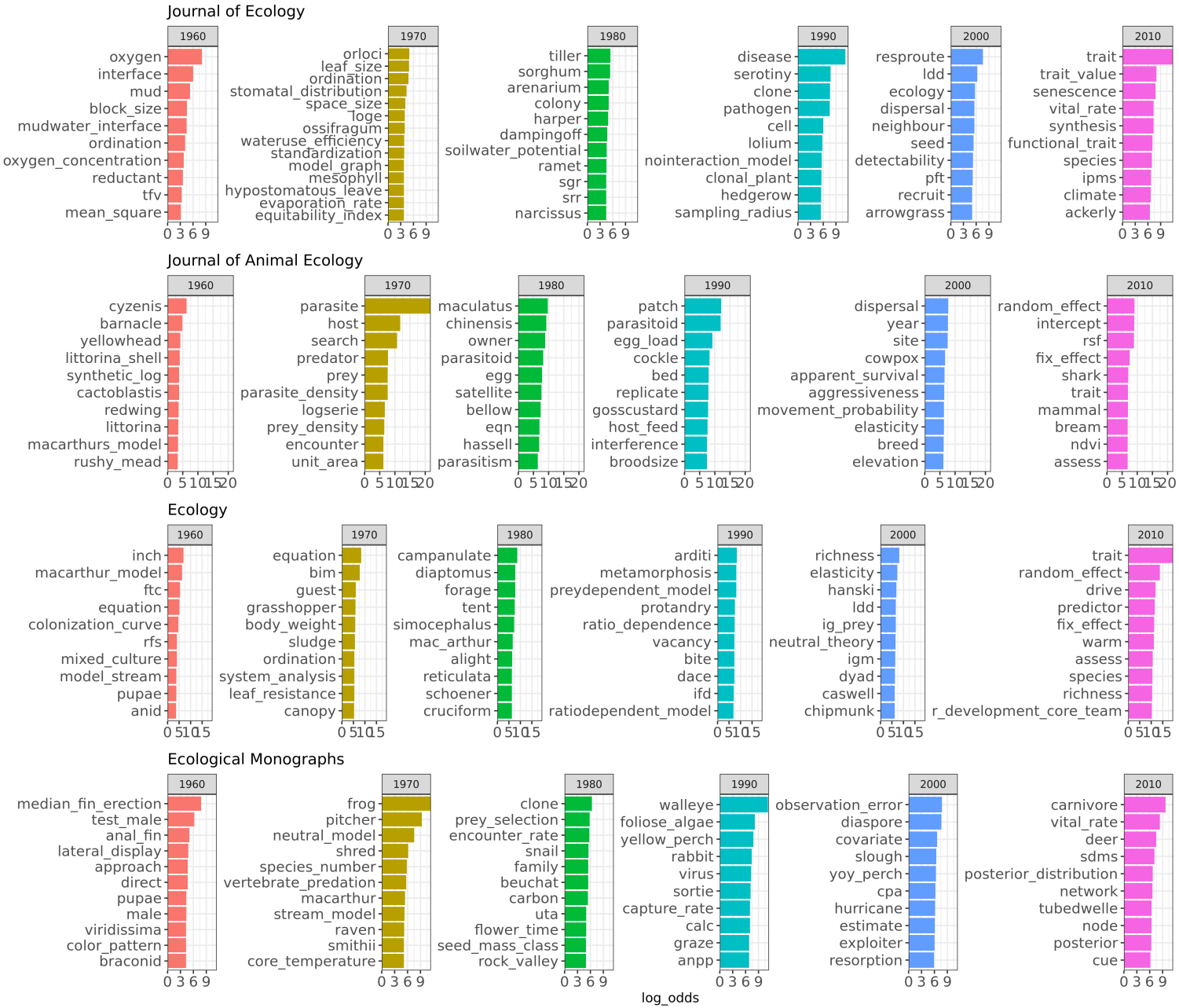


Figure 4.1: Résumé du ratio log-odds pondéré par décennie et par revue, une fois le corpus nettoyé et seulement parmi les sections contenant une occurrence de modèle.

Étant donné que l'on s'intéresse aux modèles, on peut faire de même avec les modèles que nous avons identifiés à l'aide de l'étiquetage morpho-syntaxique. On obtient les résultats suivants :

Le corpus ayant une dimension temporelle, il est possible d'examiner la manière dont la fréquence des mots a changé à travers le temps. La fréquence est calculée de la même manière que précédemment, mais on calcule la proportion par période de trois ans et pour chacune des revues. On examine les termes d'intérêt des différents types de modèles identifiés à l'aide du corpus prototypique :

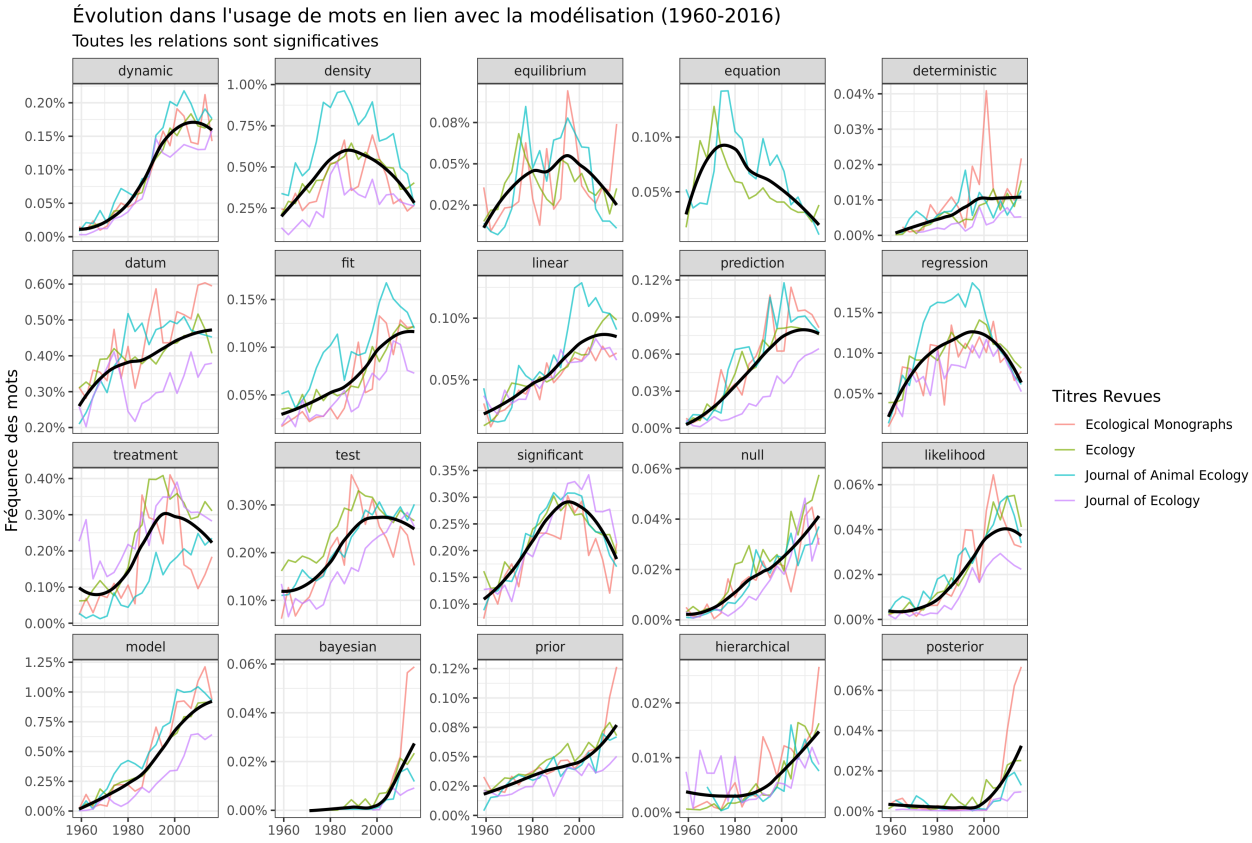


Figure 4.3: Évolution de la fréquence des mots liés à la modélisation. L'ordre des panneaux suit approximativement les différents types de modèles discutés ci-dessus, à savoir la modélisation mathématique, statistique, puis bayésienne. La ligne noire est la fréquence moyenne des quatre revues. Noter que l'axe de y diffère pour chaque panneau.

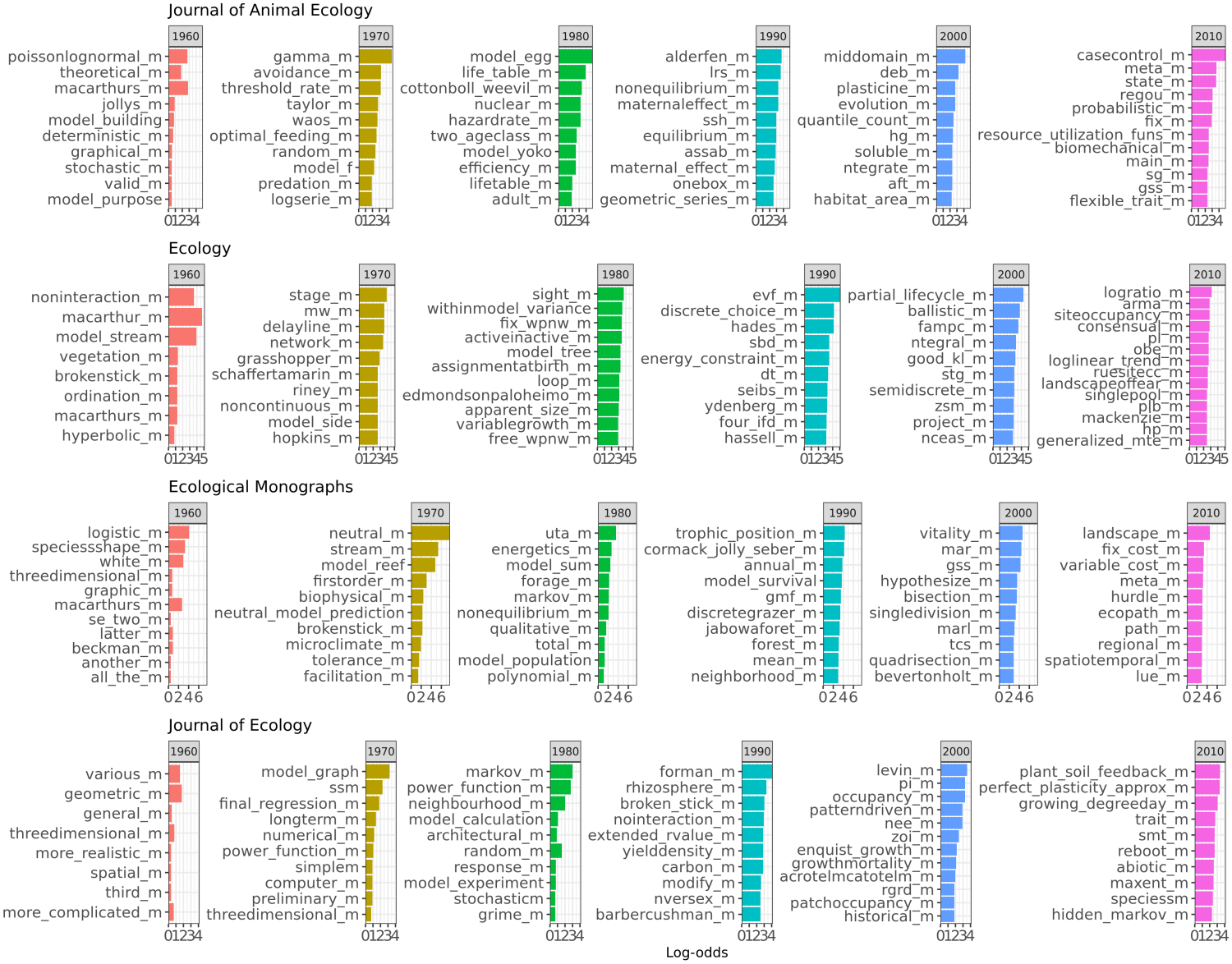


Figure 4.2: Résumé du ratio log-odds pondéré par décennie et par revue, une fois le corpus nettoyé et seulement pour les entités identifiées comme des modèles.

De loin, le terme le plus fréquent et ayant subi une plus grande augmentation est celui de « *model* ». Celui-ci commence dans les années 60 avec une fréquence de près de zéro et se termine à environ 1% du corpus dans son ensemble. Nous notons que la majorité des termes mathématiques suivent la même tendance, à savoir que leur fréquence augmente jusque dans les années 1990, puis diminuent par la suite. La fréquence des termes statistiques et bayésiens a tendance à augmenter, bien qu'il y ait quelques exceptions.

APPENDICE B: NOTES SUR LES ANALYSES SÉMANTIQUES

Détails mathématiques de la STM

On introduit les détails plus techniques de la STM pour mieux expliquer la manière par laquelle on peut informer le modèle de métadonnées. Tout comme la LDA et la CTM, la STM s'attend à avoir des documents, $d \in 1...D$, dans lesquels il y a des mots indexés par position, n 1...N. On se réfère au mot n dans le document d comme w_n, d , qui provient d'un vocabulaire d'intérêt, dénoté $v \in 1...V$. Qui plus est, nous devons spécifier un nombre de thèmes en avance, soit $k \in 1...K$. C'est ici que la CTM et la STM prennent différentes directions. On a déjà mentionné que la STM est capable d'inférer la prévalence et le contenu des sujets, c'est-à-dire dans quelle mesure un document discute de certains thèmes et la manière dont le document en discute. Pour s'y faire, la STM ajoute deux matrices de design, une pour la prévalence et une pour le contenu, dans lesquels chaque rangée est un vecteur de covariés qui contient des métadonnées sur le document. On dénote la matrice de prévalence X, laquelle est de dimension D x P, et la matrice de contenu Y, avec une dimension de D x A. Finalement, les auteurs introduisent le log de la fréquence marginale des termes du vocabulaire, m_v, à partir du compte total des mots afin de modéliser la fréquence des mots dans chaque thème en fonction des variables données.

La STM est un modèle relativement complexe qui offre plusieurs possibilités pour modéliser le processus génératif des textes. En pratique, la STM contient plusieurs modèles en un, lesquels peuvent être utilisés indépendamment les uns des autres. Dans ce qui suit, nous nous concentrons sur le modèle de prévalence et mettons de côté le modèle de contenu, car notre enquête se concentre davantage sur cet aspect que sur les différences

de contenu entre les revues. Pour mieux saisir les différences entre la LDA, la CTM, et la STM, on examine brièvement les différentes représentations graphiques de ceux-ci :

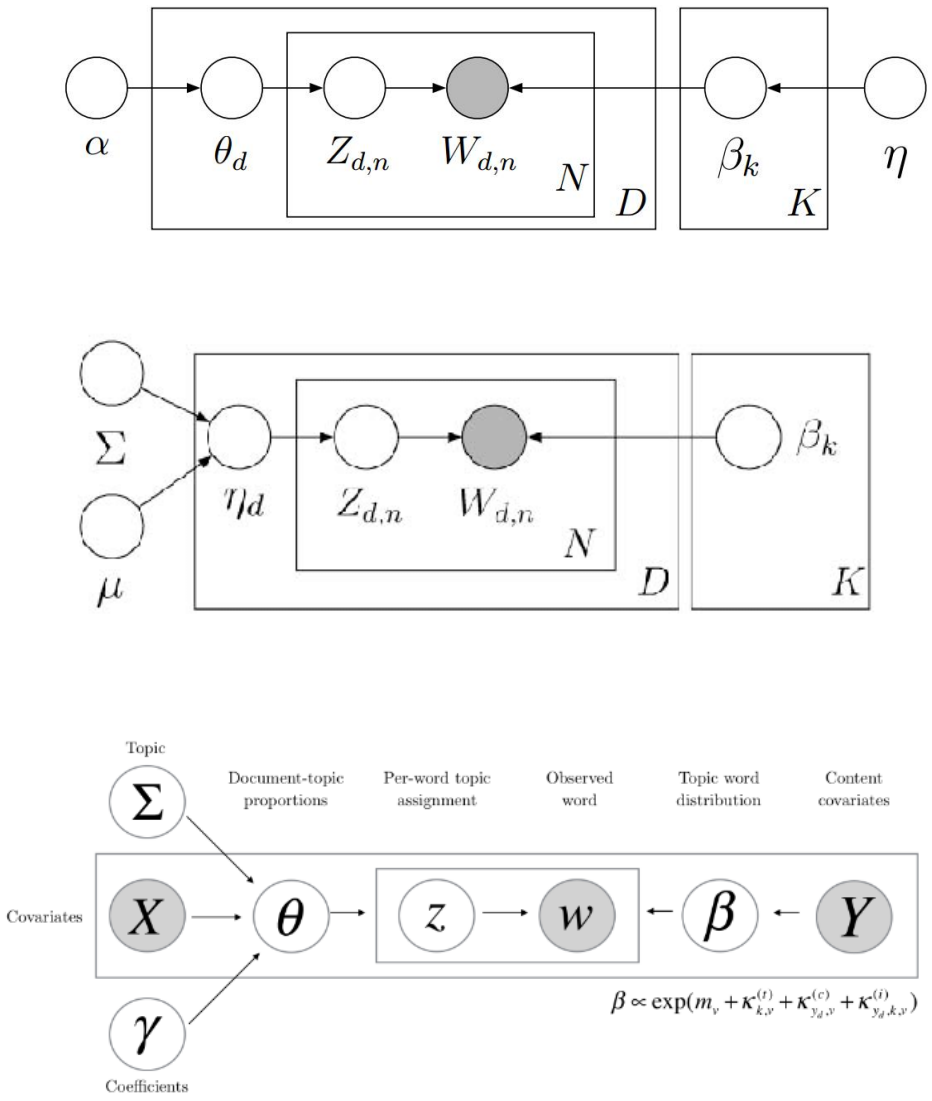


Figure 4.4: Les graphes acycliques dirigés des différents modèles. En haut, on a la LDA. Au milieu, la CTM. En bas, on a la STM. Voir description dans le corps du texte.

Ces représentations sous forme de *graphes acycliques dirigés* (DAG) permettent de facilement visualiser les dépendances conditionnelles entre les différentes composantes d'un modèle. Chaque cercle représente une variable aléatoire, et la coloration indique si la

variable est observée ou non (gris signifie que la variable est observable). Les encadrés, appelés plaques, qui sont indexés par une lettre représentent des ensembles de variables qui sont répétées. Par exemple, on peut voir que dans chacun des modèles, les mots $w_{d,n}$ et leurs distributions sont conditionnels à l'assignation à un thème, $z_{d,n}$, et le thème k. Alors que les mots et l'assignation de ceux-ci sont répétés N fois pour chaque document D dans le corpus, la distribution du thème est simplement en fonction du nombre K de thème. Le paramètre, qui représente les proportions des thèmes, se retrouve seulement à l'échelle du document.

On remarque l'ajout des matrices de design X et Y, lesquelles sont observées, dans la STM (en bas) à chaque extrémité du modèle. Ces modèles graphiques sont des objets mathématiques qui représentent comment les auteurs perçoivent leurs modèles. Ainsi, les représentations d'un même modèle peuvent varier d'un auteur à l'autre en termes de notation et de ce qui est inclus ou non, ce qui est source de difficulté pour les non-initiés. Par exemple, les auteurs de la STM laissent implicites les indexes et les plaques, et dans la CTM η_d prend le rôle de θ_d dans la LDA, alors que les auteurs de la STM préfèrent garder le θ. Lorsqu'on interprète ces graphiques, il est important de garder en tête que ceux-ci sont des descriptions incomplètes du modèle. Par exemple, les modèles graphiques ci-dessus restent muets sur les distributions de probabilité d'où proviennent les variables aléatoires. Pour un portrait plus précis, il est nécessaire d'examiner la description algébrique qui leur sont complémentaire.

On mentionne aussi la présence d'un compromis avec l'utilisation de la loi logistique-normal au lieu de la loi de Dirichlet. La loi de Dirichlet a longtemps été privilégiée, car elle est conjuguée à la loi multinomiale, qui est la loi de probabilité de notre modèle dont les paramètres sont optimisés pour connaître la proportion de mots dans les thèmes, et le mélange des thèmes dans les documents. On dit qu'une loi de probabilité est conjuguée dans le cadre bayésien lorsque la multiplication de la fonction de vraisemblance (i.e. loi multinomiale) avec le prior (i.e. loi de Dirichlet, dans le cas de la LDA) donne une pro-

babilité postérieure de la même famille que le prior. Cette qualité à des avantages computationnels non négligeables qui permet d'utiliser l'échantillonnage de Gibbs dans des délais raisonnables, même pour des corpus relativement volumineux. Avec l'utilisation de la loi logistique-normal, on perd ces avantages. Pour performer l'inférence, les auteurs de la STM ont développé un algorithme variationnel dit d'espérance-maximisation (*expectation-maximization* ; EM), lequel utilise une approximation de Laplace (Wang et Blei, 2013; Roberts *et al.*, 2019).

Détails mathématiques de l'exclusivité et de la cohérence sémantique

Concrètement, la cohérence sémantique est optimisée lorsque les mots les plus probables d'un thème co-occurrent fréquemment ensemble (Roberts, Stewart, and Tingley 2019). La cohérence C d'un thème k se calcule de la manière suivante :

$$C_k = \sum_{i=2}^{M_j} \sum_{j=1}^{i-1} \log\Big(\frac{D(v_i, v_j) + 1}{D(v_j)}\Big)$$

où $D(v_i, v_j)$ est la fréquence à laquelle deux mots co-occurrent dans un document, et M est un liste des mots les plus probables dans un thème donné.

L'indice de Fréquence-Exclusivité, ou FREX (Bischof and Airoldi 2012 ; Airoldi and Bischof 2016). L'indice FREX est une moyenne harmonique pondérée dans laquelle le rang du mot est une combinaison de sa fréquence et de son exclusivité. La formule est la suivante :

$$\text{FREX}_{k,v} = \left(\frac{\omega}{\text{ECDF}\left(\beta_{k,v} / \sum_{j=1}^{K} \beta_{j,v}\right)} + \frac{1-\omega}{\text{ECDF}\left(\beta_{k,v}\right)} \right)^{-1}$$

où k est le thème et v prend la place des mots. Nous ne rentrons pas dans les détails de la formule, mais il suffit de dire que la ECDF est la fonction de répartition empirique

appliquée à un mot dans son thème et w est un poids associé à l'exclusivité, lequel est fixé à 0.7 par les auteurs de la librairie que nous utilisons.

Diagnostics de la STM

On présente les diagnostics utilisés pour choisir le nombre de thèmes dans la modélisation thématique. On commence par un tableau qui permet de comparer le held-out likelihood, les résidus, la borne inférieure, et la cohérence sémantique :

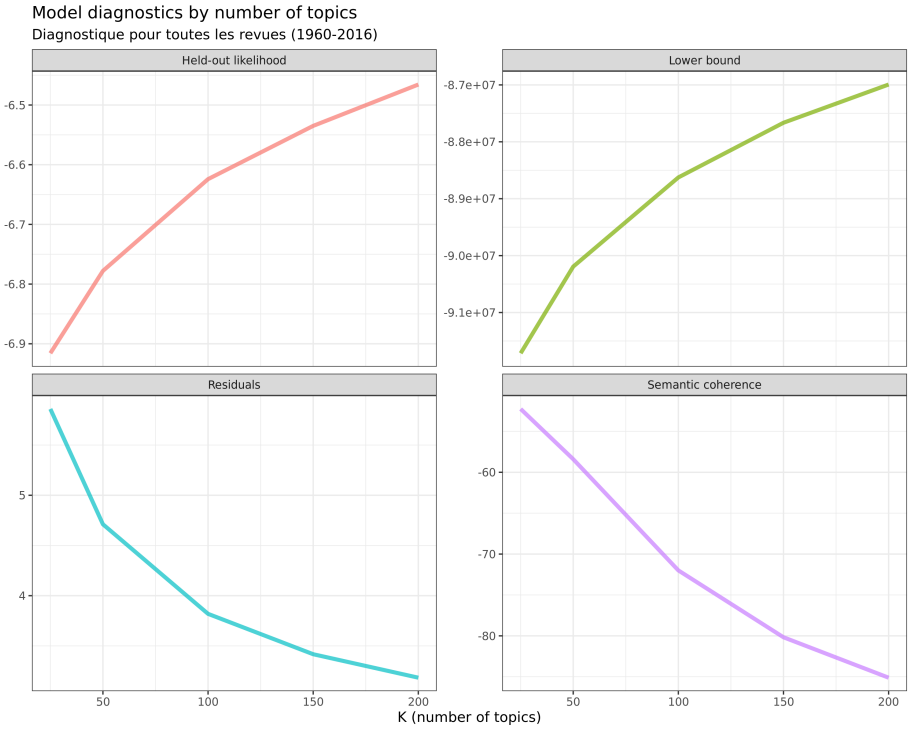


Figure 4.5: Comparaison du held-out likelihood, les résidus, la borne inférieure, et la cohérence sémantique.

Avec ce premier diagnostic, on cherche principalement à choisir le nombre de thèmes avec un held-out likelihood élevé tout en voulant minimiser les résidus. Bien qu'il n'y a pas d'optimum ni pour le held-out likehood et les résidus, on remarque que leur valeur commence à ralentir autour de 150 thèmes. De plus, si on examine le compromis entre

l'exclusivité et la cohérence sémantique :

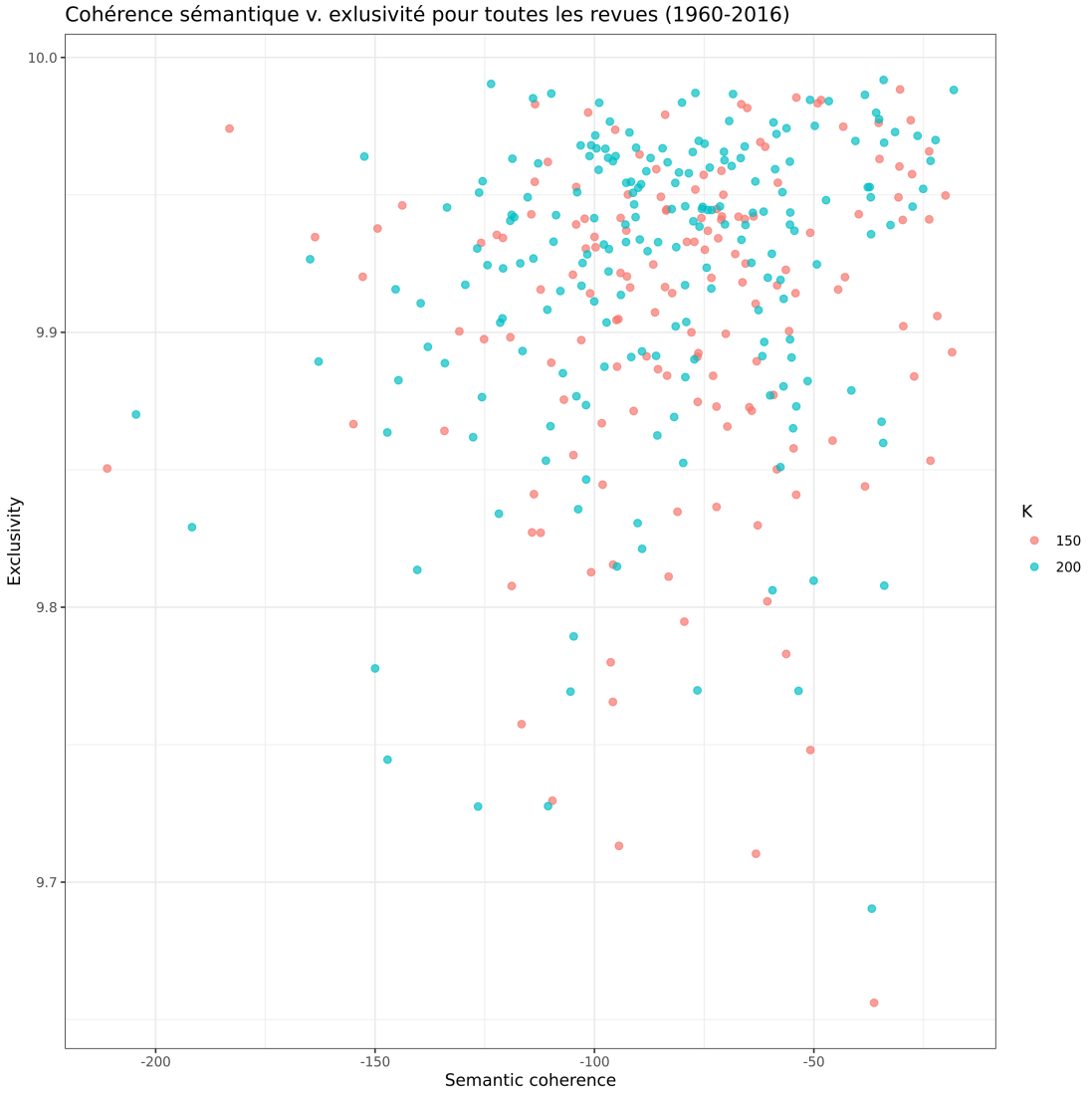


Figure 4.6: Évaluation du compromis entre exclusivité et cohérence sémantique

on remarque que l'apport sur la cohérence sémantique diminue entre 150 et 200 thèmes, i.e., on voit plusieurs thèmes dont la cohérence sémantique est plus faible à 200 qu'à 150 thèmes.

Une intuition simple pour comprendre le plongement lexical

Pour mieux comprendre l'intuition derrière le plongement lexical, on présente une intuition qui est basée sur une manière de compter les mots et la factorisation matricielle. [3] D'abord, pour construire les vecteurs, on s'intéresse aux différents contextes des mots en glissant une fenêtre sur le corpus qui comprend, disons, les quatre mots avant et après de chaque mot du corpus. En glissant cette fenêtre, on en profite pour calculer la probabilité de chaque mot et la probabilité *skipgram*, i.e., la probabilité de voir deux mots dans une même fenêtre. [4] Par la suite, on combine les deux probabilités à l'aide du *pointwise mutual information* :

$$\text{pmi}(w_i, w_j) = \log\Big(\frac{\text{skipgram}(w_i, w_j)}{\text{unigram}(w_i) \text{ x } \text{unigram}(w_j}\Big)$$

où la probabilité unigram est simplement la probabilité qu'un mot w dans le corpus. En mots, le PMI est le logarithme de la probabilité que deux mots coïncident sur la probabilité que les mots apparaissent indépendamment l'un de l'autre. Le résultat est un ratio où une valeur supérieure à 1 signifie que les deux mots cooccurrent plus souvent qu'attendu, et vice versa. À ce stade on a une matrice carrée dont la taille est le vocabulaire du corpus, et où chaque cellule représente le PMI entre deux mots. Finalement, pour avoir un espace en basse dimension, disons 100 dimensions, on performe une opération de factorisation matricielle appelée la décomposition en valeur singulière (SVD). La SVD est mieux présentée géométriquement à l'aide la visualisation suivante :

---

3. Cette intuition provient du billet de blog `https://multithreaded.stitchfix.com/blog/2017/10/18/stop-using-word2vec/` par Chris Moody.

4. La longueur de la fenêtre est un paramètre qui doit être choisi. Une grande fenêtre à tendance à capturer les relations de similarité sémantique, alors que les petites fenêtres trouvent des relations de synonymie. Dans notre analyse, on a préféré une grande fenêtre.

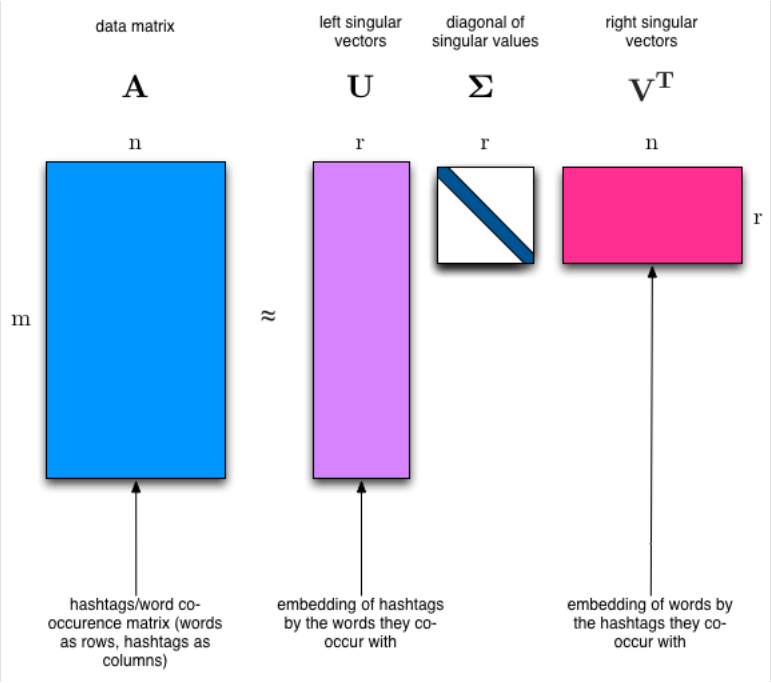


Figure 4.7: La décomposition en valeur singulière est la factorisation matricielle $A = U\Sigma V^T$.

Sans rentrer dans les détails, on trouve une approximation de rang faible d'une matrice carrée d'intérêtA, en multipliant la matrice diagonale Σ dont les coefficients sont les valeurs singulières de la matrice A, par les matrices U et V^T qui contiennent les vecteurs de mots et qui sont de taille $(n_{vocabulaire}, n_{dim})$. Géométriquement, cela revient à faire une rotation V^T suivie d'un étirement Σ le long des axes de coordonnées et d'une autre rotation U.[5] Le SVD redonne un espace orthogonal dont les "régularités linéaires" identifiées par les valeurs singulières produisent un espace vectoriel qui, dans notre exemple, encode des relations de proximité sémantiques interprétables.

5. Voir `https://en.wikipedia.org/wiki/Singular_value_decomposition` pour une animation

APPENDICE C: NOTES SUR LES TECHNIQUES DE VISUALISATION

Clustering hiérarchique agglomérative de Ward

Le clustering hiérarchique est une manière d'ordonner efficacement les valeurs d'une matrice, par exemple une matrice de corrélations, selon une mesure de similarité choisie (Holmes et Huber, 2018). Un choix de base pour la mesure de similarité est la *distance euclidienne*, i.e.,

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

où la distance entre deux vecteurs est la norme euclidienne. L'approche agglomérative, dite *bottom-up*, ordonne les valeurs en partant des valeurs les plus similaires, et de manière itérative, va agglomérer les clusters "en montant dans la hiérarchie". À chaque étape de la monté, la similarité (ou plutôt la dissimilarité) doit être recalculée entre les différentes valeurs.

La méthode de Ward quant à elle réfère à la manière dont la similarité est mesurée (Murtagh et Legendre, 2014). Plus particulièrement, la méthode de Ward vise à trouver la partition qui, à chaque itération, va minimiser la variance à l'intérieur de chaque partition, i.e., la partition qui va engendrer la plus petite augmentation de la variance à l'intérieur de chaque partition. L'implémentation utilisée est la fonction de base du langage R, et plus particulièrement la méthode WARD.D.

Il est important de mentionner que l'algorithme ne spécifie pas la granularité à privilégier dans la hiérarchie. Dit autrement, l'algorithme ne permet pas de dire si les deux

grandes branches représentent mieux le système que les embranchements au bas de la hiérarchie. Conséquemment, les couleurs utilisées dans la figure du document principal sont purement esthétiques et ne reflètent pas une propriété du clustering. Une autre limite est la rigidité des clusters. Dans notre cas, cela signifie que les topics sont ordonnés selon le premier grand cluster auquel ils sont rattachés.

UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction

UMAP est une technique de réduction de dimensionnalité basée sur la notion de variété (*manifold*) (McInnes *et al.*, 2018). La technique est similaire à l'algorithme t-SNE, lequel est hautement populaire en apprentissage machine. On explique la technique à l'aide d'un jeu de données simple intitulé les pingouins de Palmer. Le jeu de données est constitué de 334 observations de différents attributs pour trois espèces de pingouins étudiées à la station de Palmer en Antarctique. Le but de la technique est de trouver une représentation des données en deux dimensions. Pour garder l'exemple simple, on choisit seulement quatre attributs d'intérêts, à savoir la longueur et profondeur du culmen (la partie supérieur du bec de l'oiseau), la longueur des nageoires, et la masse. On visualise ce jeu de données qui habite un espace en quatre dimensions de la manière suivante :

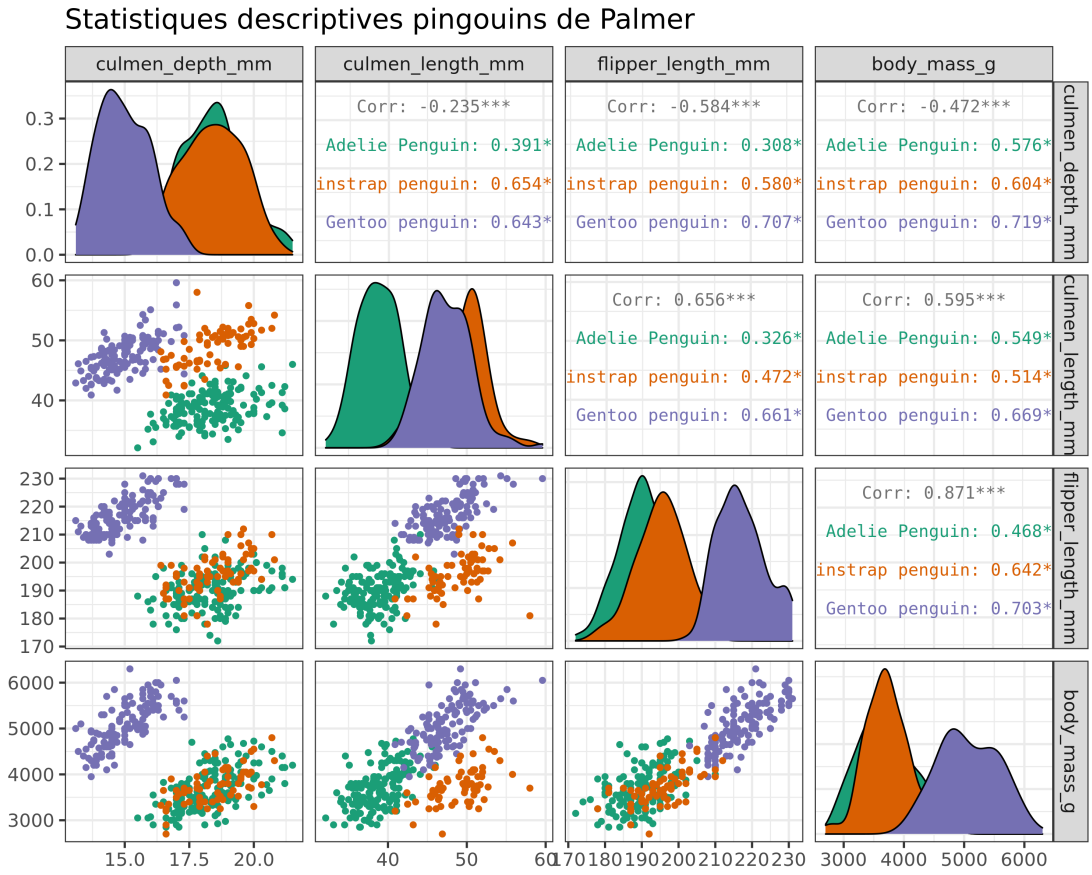


Figure 4.8: Statistiques descriptives du jeu de données

Le principal avantage de UMAP réside en sa capacité de projeter des données en haute dimension dans un espace de basse dimension, tout en préservant le plus possible la structure des données. Contrairement à l'analyse en composantes principales, qui est une projection *linéaire*, UMAP utilise les relations locales entre les points (à l'aide de complexe simplicial) pour recréer l'espace en basse dimension. La technique étant un sujet avancé en mathématique, on se contente ici de donner le résultat de cette projection pour les pingouins de Palmer, et on réfère le lecteur à l'article de Mcinnes (2018) pour les détails mathématiques et computationnels de l'algorithme. Les résultats sont les suivants :

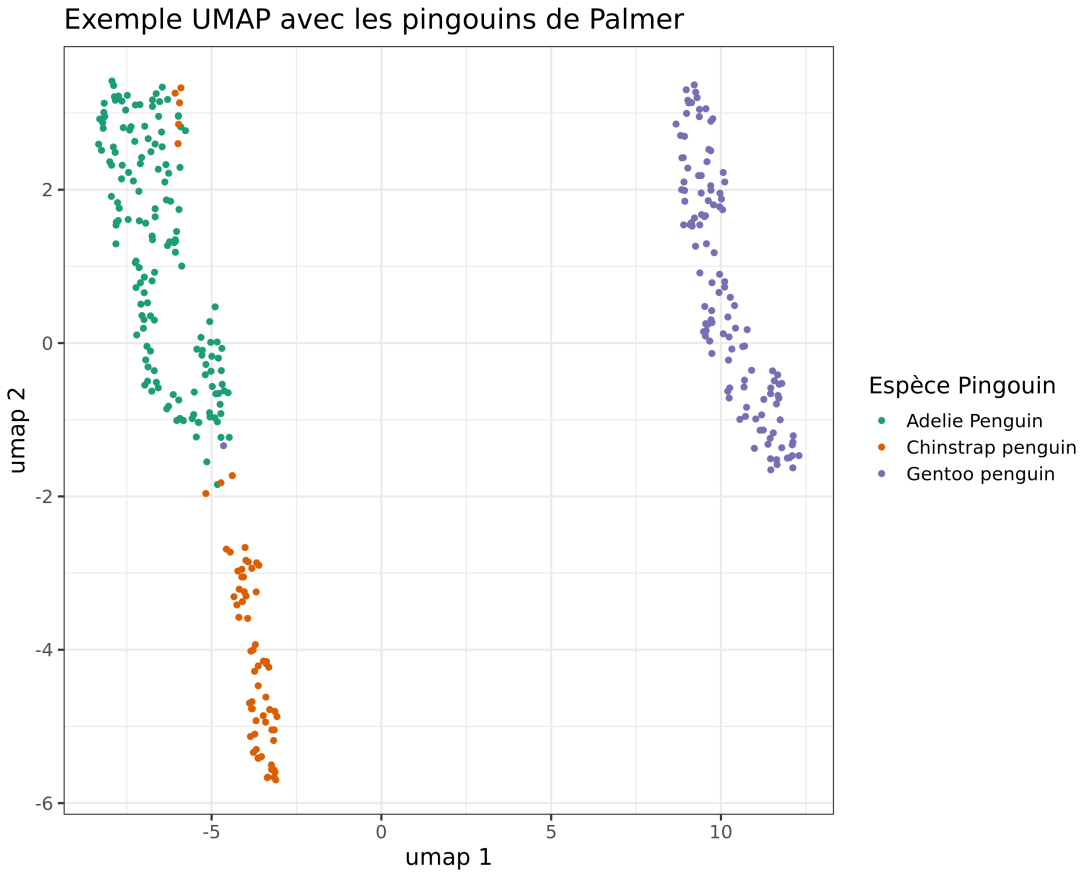


Figure 4.9: Exemple de la projection UMAP

On note que UMAP a su conserver les propriétés du jeu de données de Palmer. Bien qu'il s'agisse d'un exemple extrêmement simple, l'idée reste la même avec un jeu de données plus complexe, par exemple lorsque l'on s'intéresse aux relations de corrélation entre des topics, ou même aux relations de similarité sémantique à l'aide du plongement lexical. L'implémentation utilisée provient de la librairie UWOT en R (**?**), qui est une traduction de l'algorithme original développé par McInnes et coll. (2018).

RÉFÉRENCES

Airoldi, E. M. (2014). *Handbook of Mixed Membership Models and Their Applications* (1 éd.). Chapman and Hall/CRC.

Airoldi, E. M. et Bischof, J. M. (2016). Improving and Evaluating Topic Models and Other Models of Text. *Journal of the American Statistical Association*, *111*(516), 1381–1403.

Aitchison, J. et Shen, S. M. (1980). Logistic-Normal Distributions : Some Properties and Uses. *Biometrika*, *67*(2), 261–272.

Anderson, D. A., Burnham, K. P. et Anderson, D. R. (1998). *Model Selection and Inference : A Practical Information-Theoretic Approach*. Springer New York.

Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A. A., Leos-Barajas, V., Flemming, J. M., Nielsen, A., Petris, G. et Thomas, L. (2020). An introduction to state-space modeling of ecological time series. *arXiv :2002.02001 [q-bio, stat]*.

Bengio, Y., Ducharme, R., Vincent, P. et Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Berliner, L. M. (1996). Hierarchical Bayesian Time Series Models. Dans K. M. Hanson et R. N. Silver (dir.). *Maximum Entropy and Bayesian Methods*, Fundamental Theories of Physics, 15–22. Springer Netherlands.

Betancourt, M. (2018). Towards A Principled Bayesian Workflow.

Betancourt, M. (2019). Probabilistic Modeling and Statistical Inference.

Bischof, J. M. et Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. p. 8.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77.

Blei, D. M. et Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent Dirichlet Allocation. p. 30.

Bonvillain, N. (2010). *Cultural Anthropology*. Prentice Hall.

Boyd-Graber, J., Hu, Y. et Mimno, D. (2017). *Applications of Topic Models*.

Crossley, N., Bellotti, E., Edwards, G., Everett, M. G., Koskinen, J. et Tranmer, M. (2015). *Social Network Analysis for Ego-Nets*. SAGE Publications Ltd.

Diaconis, P. (2008). The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, *46*(2), 179–205.

Evans, M. R., Grimm, V., Johst, K., Knuuttila, T., de Langhe, R., Lessells, C. M., Merz, M., O'Malley, M. A., Orzack, S. H., Weisberg, M., Wilkinson, D. J., Wolkenhauer, O. et Benton, T. G. (2013). Do simple models lead to generality in ecology ? *Trends in Ecology & Evolution*, *28*(10), 578–583.

Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-1955. In *Studies in Linguistic Analysis. Special volume of the Philological Society*. Oxford, UK : Basil Blackwell.

Fraassen, B. C. v. (1980). *The Scientific Image*. Clarendon Press.

Frigg, R. et Hartmann, S. (2020). Models in Science. In E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, (spring 2020 éd.).

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. et Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, *182*(2), 389–402.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. et Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

Gelman, A., Hill, J. et Vehtari, A. (2020). *Regression and Other Stories*. Cambridge University Press.

Gelman, A., Simpson, D. et Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, *19*(10), 555.

Grimmer, J. et Stewart, B. M. (2013). Text as Data : The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297.

Hall, D., Jurafsky, D. et Manning, C. D. (2008). Studying the history of ideas using topic models. Dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, p. 363., Honolulu, Hawaii. Association for

Computational Linguistics.

Hamilton, W. L., Leskovec, J. et Jurafsky, D. (2018). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv :1605.09096 [cs]*. arXiv : 1605.09096.

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2-3), 146–162.

Hesse, M. (1966). *Models and analogies in science*. South Bend, IN : University of Notre Dame Press.

Hilborn, R. et Mangel, M. (1997). *The Ecological Detective : Confronting Models with Data*. Princeton University Press.

Hobbs, T. N. et Hooten, M. B. (2015). *Bayesian models a statistical primer for ecologists*. Princeton University Press.

Hodges, J. S. (2019). Statistical methods research done as science rather than mathematics. *arXiv :1905.08381 [stat]*. arXiv : 1905.08381.

Holmes, S. et Huber, W. (2018). *Modern Statistics for Modern Biology*. Cambridge University Press.

Jordan, M. I. (2004). Graphical Models. *Statistical Science*, *19*(1), 140–155.

Jurafsky, D. et Martin, J. M. (2019). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

Kroese, D. P., Brereton, T., Taimre, T. et Botev, Z. I. (2014). Why the Monte Carlo method is so important today : Why the MCM is so important today. *Wiley Interdisciplinary Reviews : Computational Statistics*, *6*(6), 386–392.

Kruschke, J. K. (2010). *Doing Bayesian Data Analysis : A Tutorial with R and BUGS*. Academic Press.

Levins, R. (1966). The Strategy of Model building in Population Biology. *American Scientist*, *54*(4), 421–431.

Lloyd, E. A. (1988). The Semantic Approach and Its Application to Evolutionary Theory. *PSA : Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1988*(2), 278–285.

Malaterre, C., Chartier, J.-F. et Pulizzotto, D. (2019). What Is This Thing Called *Philosophy of Science* ? A Computational Topic-Modeling Perspective, 1934–2015. *HOPOS : The Journal of the International Society for the History of Philosophy of Science*, *9*(2), 215–249.

Matthewson, J. et Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, *170*(1), 169–190.

McElreath, R. (2020). *Statistical Rethinking : A Bayesian Course with Examples in R and Stan* (2 éd.). Chapman and Hall/CRC.

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D. et Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, *41*(6), 607–625.

McInnes, L., Healy, J. et Melville, J. (2018). UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv :1802.03426 [cs, stat]*. arXiv : 1802.03426.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. p. 9.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. et McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. Dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 262–272., Stroudsburg, PA, USA. Association for Computational Linguistics. event-place : Edinburgh, United Kingdom.

Murtagh, F. et Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward's Criterion ? *Journal of Classification*, *31*(3), 274–295.

Nitecki, M. H. et Hoffman, A. (dir.) (1987). *Neutral models in biology*. New York : Oxford University Press.

Odenbaugh, J. (2007). The strategy of "The strategy of model building in population biology". *Biology & Philosophy*, *21*(5), 607–621.

Otto, S. P. et Day, T. (2007). *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*. Princeton University Press.

Patterson, T., Thomas, L., Wilcox, C., Ovaskainen, O. et Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in Ecology & Evolution*, *23*(2), 87–94.

Railsback, S. F. et Grimm, V. (2012). *Agent-Based and Individual-Based Modeling : A Practical Introduction*. Princeton University Press.

Roberts, M. E., Stewart, B. M. et Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical*

Association, *111*(515), 988–1003.

Roberts, M. E., Stewart, B. M. et Tingley, D. (2019). stm : An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. et Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Roughgarden, J. (1998). *Primer of Ecological Theory*. Prentice Hall.

Spiegelhalter, D. (2019). *The Art of Statistics : Learning from Data*. Penguin UK.

Suppes, P. (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese*, *12*(2/3), 287–301.

Tversky, A. (1977). Features of Similarity. *Psychological Review*, *84*(4), 327–352.

Vehtari, A., Gelman, A. et Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. arXiv : 1507.04544.

Wallach, H. M., Murray, I., Salakhutdinov, R. et Mimno, D. (2009). Evaluation methods for topic models. Dans *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, 1–8., Montreal, Quebec, Canada. ACM Press.

Wang, C. et Blei, D. (2013). Variational Inference in Nonconjugate Models. *Journal of Machine Learning Research*, *14*, 1005–1031.

Weisberg, M. (2003). *WHEN LESS IS MORE : TRADEOFFS AND IDEALIZATION IN MODEL BUILDING*. (Thèse de doctorat).

Weisberg, M. (2006). Robustness Analysis. *Departmental Papers*, *4*, 15.

Weisberg, M. (2007). Forty Years of 'The Strategy' : Levins on Model Building and Idealization. *Biology & Philosophy*, *21*(5), 623–645.

Weisberg, M. (2013). *Simulation and Similarity : Using Models to Understand the World*. OUP USA.

Weisberg, M. et Muldoon, R. (2009). Epistemic Landscapes and the Division of Cognitive Labor*. *Philosophy of Science*, *76*(2), 225–252.

Weisberg, M. et Reisman, K. (2008). The Robust Volterra Principle*. *Philosophy of Science*, *75*(1), 106–131.